Identifying Task Groupings for Multi-Task Learning Using Pointwise $\mathcal{V}$-Usable Information

Yingya Li [1] **Timothy Miller**[1] **Steven Bethard**[2] **Guergana Savova**[1]

[1]Boston Children's Hospital and Harvard Medical School, {firstname.lastname}@childrens.harvard.edu

[2] University of Arizona, bethard@email.arizona.edu

Abstract

The success of multi-task learning can depend heavily on which tasks are grouped together. Naively grouping all tasks or a random set of tasks can result in negative transfer, with the multi-task models performing worse than single-task models. Though many efforts have been made to identify task groupings and to measure the relatedness among different tasks, it remains a challenging research topic to define a metric to identify the best task grouping out of a pool of many potential task combinations. We propose a metric of task relatedness based on task difficulty measured by pointwise $\mathcal{V}$-usable information (PVI). PVI is a recently proposed metric to estimate how much usable information a dataset contains given a model. We hypothesize that tasks with not statistically different PVI estimates are similar enough to benefit from the joint learning process. We conduct comprehensive experiments to evaluate the feasibility of this metric for task grouping on 15 NLP datasets in the general, biomedical, and clinical domains. We compare the results of the joint learners against single learners, existing baseline methods, and recent large language models, including Llama 2 and GPT-4. The results show that by grouping tasks with similar PVI estimates, the joint learners yielded competitive results with fewer total parameters, with consistent performance across domains.

1 Introduction

Multi-task learning (MTL) learns shared representations across tasks and jointly optimizes the losses of all included tasks, which reduces the risk of over-fitting (Caruana, 1997). Compared to single-task learning (STL), MTL has been shown to improve performance and generalization capabilities in many natural language processing (NLP) tasks (e.g. Liu et al., 2019a; Peng et al., 2020; Khandelwal and Britto, 2020; Zhang et al., 2020a). However, empirical results also suggest that MTL is not always effective and naively grouping tasks brings negative transfer (Wu et al., 2019; Alonso and Plank, 2017; Wu et al., 2019; Guo et al., 2020; Fifty et al., 2021; Aghajanyan et al., 2021; Li et al., 2023a). The space of possible task combinations can be massive, and naively searching that space to find the best joint learning models is not efficient (Fifty et al., 2021; Song et al., 2022).

To find the best task combinations, some recent studies have developed new optimization methods that focus on measuring the relatedness among tasks (Fifty et al., 2021; Song et al., 2022; Ni et al., 2023; Li et al., 2023a; Ni et al., 2023; Li et al., 2023b). For example, Vu et al. (2020) applied task embedding to predict the transferability of source tasks to a target task. Fifty et al. (2021) compared the inter-task affinity by examining how one task's gradient updates on the shared parameters would influence the objective of another task. Song et al. (2022) leveraged a meta-learning framework on task combinations. Li et al. (2023a) applied surrogate models to identify negative transfers among different groupings during MTL and to identify the best combinations for joint learning. However, finding the optimal task grouping usually involves combining many, if not all, tasks for training and optimization, which becomes computationally intensive as the number of tasks increases. For a deeper understanding of the task-relatedness of MTL in neural networks, researchers also provide initial clues to formalize the definition through measurable variables. Specifically, some work suggests that auxiliary tasks with compact and more uniform label distributions are preferable for semantic sequence prediction problems (Alonso and Plank, 2017). Others found that gains are more likely to occur for main tasks that plateau quickly with non-plateauing auxiliary tasks (Bingel and Søgaard, 2017). In certain domains like financial NLP tasks, study results show that MTL works well when tasks are related and with diverse skills (Ni et al., 2023).

Manuscript submitted to ACL conferences.

Nevertheless, we still lack a shared definition of task-relatedness or a metric to measure the amount of cross-task usable information for a given model under the joint learning context.

This work studies the use of pointwise $\mathcal{V}$-usable information (PVI) (Ethayarajh et al., 2022) to measure the usable information of different datasets and to jointly train tasks with similar information gains given a model. PVI, recently introduced by Ethayarajh et al. (2022), estimates the difficulty of data instances for a given model in supervised learning. It builds on the predictive $\mathcal{V}$-information framework (Xu et al., 2020) which incorporates mutual information and the coefficient of determination to quantify data instance difficulty. The metric applies instance-level predictions to quantify how much information a given model can extract from a dataset. The higher the PVI estimate, the easier it is for the model to represent a given data point. Under this context, we cast PVI as an estimate of task-relatedness to guide MTL. By grouping tasks according to similar PVI distributions, or in other words, tasks of comparable difficulty, we hypothesize that this approach promotes model generalization across the targeted tasks in MTL.

To investigate the feasibility of identifying the best task groupings for MTL using PVI, we conducted experiments with 15 NLP datasets in the general, biomedical and clinical domains. We compared the MTL results with task groupings selected by the PVI estimate distributions against the best-performing fine-tuned single-learner models. The performances were also compared against recent large language models (LLMs), including Llama 2 (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023) which have demonstrated their ability as general-purpose NLP task solvers across a wide range of NLP tasks, either with or without downstream data adaptation (Liu et al., 2023; Brown et al., 2020; Chowdhery et al., 2023; Qin et al., 2023). We also provide comparison to two baseline task grouping methods: task embedding (Vu et al., 2020) and surrogate models (Li et al., 2023a) considered state of the art. Our contributions are summarized as follows:

- We introduce a new method to identify task groupings using PVI estimates (Ethayarajh et al., 2022) for MTL.
- We provide a thorough empirical analysis across NLP tasks in different domains, demonstrating that our method could effectively find high-performing task groupings that achieve or surpass STL performance.
- We offer a new perspective on using PVI estimates for task groupings in MTL highlighting its effectiveness for tasks that fall in roughly the same domain.

2 Related Work

2.1 MTL with similar tasks

Jointly training similar tasks is the main premise of MTL especially in the era of neural models and transfer learning. The more similar shared tasks are, the more hidden units would be shared in a given model, which would potentially benefit the joint training process through these shared representations (Caruana, 1997). In the context of MTL, measuring task similarity as well as automatically and reliably determining the optimal task grouping from numerous possible configurations remains challenging. Zhang et al. (2023) surveyed the efforts in the NLP field for task relatedness and training methods. Overall, the empirical selection of similar tasks remains the most commonly used method, and in most cases, the problem of deciding which tasks to combine for MTL is often left to human experts (Zhang and Yang, 2021). In recent years, a few methods have been developed to automatically select similar tasks (Vu et al., 2020; Fifty et al., 2021; Aribandi et al., 2021; Song et al., 2022; Ni et al., 2023; Li et al., 2023a; Ni et al., 2023; Li et al., 2023b), though a shared definition of task similarity for joint training is still lacking (Zhang et al., 2023). As a result, negative transfers among tasks in joint training have been observed by researchers (Wu et al., 2019; Alonso and Plank, 2017; Wu et al., 2019; Guo et al., 2020; Fifty et al., 2021; Aghajanyan et al., 2021; Li et al., 2023a). It would also be computationally costly to search for the best task grouping by iteratively combining all tasks in pairs or n-task groupings. Therefore, a straightforward and efficient method is needed to identify the best task groupings for MTL.

2.2 PVI and dataset difficulty

Understanding the difficulty of a task helps guiding the machine learning process, e.g. what architectures and classification methods are feasible (Torralba and Efros, 2011; Zhao et al., 2022; Cui et al., 2023). The significance of dataset difficulty has been widely discussed in the NLP field (Hahn et al., 2021; Perez et al., 2021; Zhao et al., 2022; Gadre

et al., 2023). As an extension of the predictive $\mathcal{V}$-information framework (Xu et al., 2020), PVI measures dataset or instance difficulty by the lack of usable information given a model. Algorithm 1 shows how PVI score (Ethayarajh et al., 2022) is calculated (see Section 3 for a detailed description). A high PVI estimate indicates a good representation of the input in the model, and thus the instance is regarded to be easier for the given model. Contrarily, a low PVI estimate indicates the input has less usable information to the model, and the instance is thus considered to be harder. Under the supervised learning context, PVI offers a practical metric that is able to compare the dataset difficulties among different NLP tasks.

PVI has proven effective across downstream tasks, such as quality estimate for universal dependencies (Kulmizev and Nivre, 2023), informativeness evaluation in reasoning chain (Prasad et al., 2023) and data augmentation for intent detection (Lin et al., 2023). Lu et al. (2023) further adapted PVI to in-context learning for LLMs, illustrating its consistency and effectiveness in this new paradigm. In the current study, we propose to explore this metric as a proxy for task similarity to identify the best task groupings for MTL.

3 Method

In this section, we describe our proposed method of utilizing PVI for task groupings in MTL. The method consists of two stages: 1) calculating the PVI estimates for each task; 2) grouping tasks based on PVI score distributions for MTL.

Algorithm 1 The calculation of PVI estimate

1: **Input:** a dataset $\mathcal{D}$, a model $\mathcal{G}$, a test instance $(x, y) \notin \mathcal{D}$
2: $g' \leftarrow$ fine-tune $\mathcal{G}$ on $\mathcal{D}$
3: $\varnothing \leftarrow$ empty string
4: $g \leftarrow$ fine-tune $\mathcal{G}$ on $\{(\varnothing, y_i)|(x_i, y_i) \in \mathcal{D}\}$
5: $\text{PVI}(x \rightarrow y) \leftarrow -\log_2 g[\varnothing](y) + \log_2 g'[x](y)$

Algorithm 1 shows how PVI score (Ethayarajh et al., 2022) is calculated. Specifically, it involves fine-tuning a given model $\mathcal{G}$ in two different and separate setups indicated by g' and g. For g', $\mathcal{G}$ is fine-tuned with the input-target pairs $\{(x_i, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For g, $\mathcal{G}$ is fine-tuned only using the targets, $\{(\varnothing, y_i)|(x_i, y_i) \in \mathcal{D}\}$. For each instance in a dataset $\mathcal{D}$, given a model $\mathcal{G}$, a higher PVI score indicates that the instance x provides more usable information to the model $\mathcal{G}$.

Depending on the individual PVI estimates for each instance given a model $\mathcal{G}$, our proposed method further groups the datasets by their PVI score distributions. The similarities of distributions over the instances of the datasets are compared using either a paired t-test or a one-way analysis of variance (ANOVA), depending on the number of datasets included with each group.

Datasets with similar PVI score distributions (i.e. not statistically significantly different) are jointly trained in MTL with a hard parameter-sharing strategy, within which the lower layers of a neural network model learn representations across all tasks grouped together, and on top of the shared layers, there are task-specific layers that learn the representations specific to each task (shown in Figure 1). We used a shared encoder that processes inputs from all tasks and task-specific output heads on top of the shared encoder. Under this setting, all the lower layer parameters are shared among all tasks, thus the total number of parameters of the joint model is lower than when one model is trained for each task. In addition, parameter sharing helps in preventing overfitting, especially when the amount of data for each task is limited.

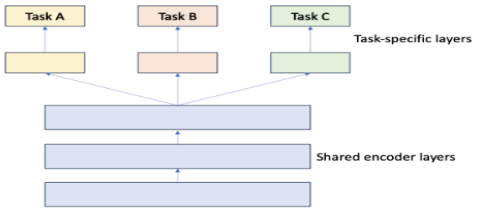

Figure 1: An illustration of the architecture for hard parameter sharing of hidden layers in MTL.

4 Experimental Setup

4.1 Datasets and tasks

We conducted experiments on 15 datasets, including 7 general, 2 biomedical, and 6 clinical NLP benchmarks. The details of the datasets, including the data statistics of train, dev, and test splits are shown in Tables 9 and 10 in Appendix A.1.

General-domain tasks The general domain tasks include BoolQ (Clark et al., 2019), CB (De Marneffe et al., 2019), RTE (Wang et al., 2019), COPA (Roemmele et al., 2011), WiC (Pilehvar and Camacho-Collados, 2019), CoLA (Warstadt et al., 2019), SST2 (Socher et al., 2013). They are tasks from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019), commonly used as benchmarks

for natural language understanding covering question answering, natural language inference, word sense disambiguation, linguistic acceptability, and sentiment analysis.

Biomedical tasks We used the Health Advice (Li et al., 2021) and Causal Language (Yu et al., 2019) datasets representing two fundamental tasks in the biomedical domain: classification and reasoning. Health Advice categorizes biomedical literature sentences into "no advice", "weak advice", and "strong advice" based on recommendation strength. Causal Language identifies correlational and causal statements in biomedical research findings, labeling each as "correlation", "conditional causation", "direct causation" or "no relationship".

Clinical tasks The clinical datasets cover 6 tasks from the THYME corpus (Temporal Histories of Your Medical Events) (Styler IV et al., 2014), SHARP Seed (Seed), and SHARP Stratified (Strat). The THYME corpus contains 594 clinical and pathology notes on colon cancer patients, annotated with different syntactic and semantic information. We used the THYME+ version of the corpus (Wright-Bettner et al., 2020) and worked with two tasks: negation (THYMENeg) and contextual modality (THYMEMod). Negation indicates if an entity or an event is "negated" or "not negated". Contextual modality flags an entity or an event as "actual" (an event has happened or is scheduled to happen), "hedged/uncertain" (an event is mentioned with some degree of hedging/uncertainty), "hypothetical" (an event is conditional on some other event, usually introduced by "if") or "generic" (an event is mentioned in a general sense, not related to a specific person). The Seed corpus contains notes from patients with pulmonary arterial disease from the Mayo Clinic and patients with breast cancer from Seattle Group Health Cooperative (now part of Keiser Permanante). The Strat corpus contains notes from patients from the Mayo Clinic, with a variety of specialties and note types representing the entire Electronic Medical Records (EMR). We selected the tasks of Negation (SeedNeg, StratNeg) and Uncertainty (SeedUncert, StratUncert) from the two datasets. All of these clinical corpora are de-identified.

4.2 Models

For the general and biomedical datasets, the base settings for STL and MTL were based on the setting in Liu et al. (2019b) using roberta-large as the pre-trained model. Though prior work suggests a modest benefit from domain-specific pre-training of the two biomedical tasks (Yu et al., 2019; Li et al., 2021), we used one pre-trained model across the general and biomedical tasks, to allow task combinations across these domains.

Given the unique characteristics of clinical language which is the EMR clinical narrative, we used Bio+Clinical BERT (Alsentzer et al., 2019) as the pre-trained model for STL and MTL for the 6 selected clinical tasks. The model was trained on all clinical notes from the MIMIC III dataset (Johnson et al., 2016) atop of BioBERT (Lee et al., 2020) and has been shown to work better for clinical NLP tasks in comparison to other encoder-only BERT-based models (Alsentzer et al., 2019).

We compared our approach with two baseline task grouping methods: task embedding (Vu et al., 2020) and surrogate models (Li et al., 2023a). The task embedding method effectively identifies transferability across 33 NLP tasks (Vu et al., 2020). The surrogate model method (Li et al., 2023a) outperforms other existing task grouping strategies in both NLP and computer vision tasks, such as HOA (Standley et al., 2020) and TAG (Fifty et al., 2021).

To further evaluate the MTL performance using PVI estimates, we compared the results with recent LLMs including LlaMA2-7B-chat, LlaMA2-70B-chat (Touvron et al., 2023) and GPT-4 (Achiam et al., 2023), using few-shot prompting with the number of shots equal to the number of classes in each task. This comparison aims to determine if LLMs' performance could match or exceed our best MTL models, as LLMs have been claimed to handle various NLP tasks with impressive performance even without downstream adaptation, and thus could be considered a type of multi-task learners. Tables 17 and 18 in Appendix A.2 show examples of prompts used in the study. We limited the evaluation of GPT-4 to experiments on the 7 general and 2 biomedical datasets, due to the HIPAA (Health Insurance Portability and Accountability Act) regulation of the clinical datasets.[1]

4.3 Setting and evaluation

To obtain the PVI estimates for all datasets, we fine-tuned the base models for 10 epochs with a batch size of 32, and a learning rate of $2e$-5 through HuggingFace's Transformer API (Wolf et al., 2020).

[1]We do not have access to the HIPAA-compliant GPT-4 and are not allowed to pass the data to the public GPT-4 API.

Unless specified, all other settings in our experiments are the default ones.

For STL and MTL with different task groupings, we fine-tuned the models with different seed values, epoch numbers, and learning rates. More details of the experiments are in Appendix A.1. In consideration of the skewed label distribution for some of the tasks, we adopted both accuracy and macro F1-score as our evaluation metrics. All the datasets were separated into train, dev and test sets. Results for all tasks, using different models and settings are reported on the test set.[2]

5 Results

We conducted paired t-tests to compare the PVI estimates of pairs of tasks (group size = 2) using the same model. This analysis yielded 36 task groupings for the general and biomedical domains and 15 groupings for the clinical domain. Table 1 shows the results of task groupings where there is no statistically significant difference in PVI estimates between the grouped tasks (p-value > 0.01) and therefore the two tasks have a similar level of difficulty for the given model.

Task Group	T-statistic	P-value
Health Advice-Causal Language	1.398	0.162
Causal-CB	2.411	0.016
BoolQ-RTE	1.775	0.076
CB-COPA	1.328	0.185
CB-CoLA	1.237	0.216
CB-SST2	-1.660	0.097
WiC-COPA	-2.358	0.019
COPA-CoLA	-0.072	0.943
THYMENeg-SeedNeg	1.680	0.093
THYMEMod-SeedNeg	-1.740	0.082
THYMEMod-StratNeg	0.585	0.559
SeedNeg-StratNeg	2.546	0.011
SeedUncert-StratUncert	0.380	0.704

Table 1: Task groupings (group size = 2) with PVI estimates that are not significantly different based on paired t-test results (p-value > 0.01).

For task groupings with more than 2 tasks (group size > 2), we used one-way ANOVA to examine the difference of PVI estimates across different tasks. Table 2 shows the results of task groupings with similar PVI estimates distribution (p-value > 0.01). We also observed that in group sizes larger than 3, no tasks have similar difficulty levels, as measured by the one-way ANOVA of PVI estimates. This means that among all tasks we experimented with, no more than three tasks are equally challenging for the pre-trained model, i.e. no more than three tasks would fit our criteria for task grouping.

Task Group	F-statistic	P-value
CB-CoLA-COPA	0.813	0.444
THYMEMod-SeedUncert-StratNeg	4.330	0.013
THYMEMod-SeedNeg-StratNeg	1.801	0.163

Table 2: Task groupings (group size > 2) with PVI estimates that are not significantly different based on one-way ANOVA results (p-value > 0.01).

5.1 MTL results of grouped tasks with similar PVI estimates

To test the feasibility of utilizing task difficulty for task grouping, we jointly trained on the datasets with similar PVI estimates. Tables 3 and 4 show the accuracy and macro F1 results of the single-task learners and joint learning models when combining tasks with similar estimates.[3] For the tasks that have the PVI estimate similar to more than one task, such as CB as shown in Table 1, we selected the grouping with the larger T-statistic values for the joint training.

For all general and biomedical tasks the joint learning (MTL) results when grouping two tasks with statistically similar PVI estimates are either similar or better than the STL (margin of error factored in). Based on the accuracy and F1 scores, for the 9 general and biomedical tasks, MTL for groupings (group size = 2) of statistically similar PVI estimates produced results that were either comparable to or better than those of the single learners for all datasets. For WiC, Health Advice, and CoLA when paired with COPA, Causal Language, CB and COPA the MTL results are within the margin of error when compared to the STL results. The CB task has the largest F1-score improvement ($\Delta = 0.257$). Similar patterns are also observed with grouping sizes larger than 2 when 3 tasks with similar PVI estimates are combined for joint training.

Results on the clinical tasks are in Table 4. The observed consistent improvement of PVI-based

[2]GLUE and SuperGLUE tasks have no ground truth labels in the original test set. We divided the original training set into a train and a dev split. Like prior studies (Zhang et al., 2020b; Du et al., 2022), the original dev set serves as the held-out, eyes-off dataset, which was used only to report the results in this paper.

[3]In MTL, datasets are jointly trained in pairs (group size = 2) or sets of three (group size = 3). For example, BoolQ and RTE in the first row were jointly trained. MTL scores (accuracy and macro-F1) indicate performance for each dataset in a group. Average results were reported with a 95% confidence interval from 5 random seed samples (t = 2.776).

MTL grouping as compared to single learners may be due to the more constrained and formulaic nature of the language in this field, which better supports task semantics than the broader variability seen in the general domain.

The same trends are exhibited when we experimented with other language models (bert-base-uncased, BioBERT) shown in Tables 15-16 in Appendix A.1.

Compared with the best MTL learning results of the two baseline approaches, shown in Tables 11-14 in Appendix A.1, the method of utilizing similar PVI estimates also shows either the same (e.g., Health advice and Causal Language) or a large improvement among the tasks in different domains. Results for THYMENeg combined with THYMEMod are similar within the margin of error. Our method trends toward better performance when compared to the recent surrogate model method (Li et al., 2023a) reported to outperform other existing task grouping strategies on both NLP and computer vision tasks, such as HOA (Standley et al., 2020) and TAG (Fifty et al., 2021).

5.2 MTL results of grouped tasks with different PVI estimates

To further examine how task difficulty may affect the performance of joint learning when grouping tasks by their PVI estimates, we combined tasks with the most statistically different PVI estimates (p-value < 0.01) for joint learning. Tables 5 and 6 show the results on the general, biomedical, and clinical tasks individually. Compared to STL and MTL with task grouping of similar PVI estimates, the performance is much lower. Even though the CB dataset showed a better performance when combined with the Health Advice datasets, it came at a cost for the other task - the performance of Health Advice was lower compared to its single learner result. On the other hand, the results of the Health Advice dataset are similar to those from MTL with similar PVI grouping however it comes at a cost for the the other task (as shown in Table 5) .

We observed a similar pattern for the tasks in the clinical domain, except for the slight improvement over STL when THYMEMod is jointly trained with THYMENeg, and a better performance of StratUncert when combined with THYMEMod (at cost for THYMEMod). This result concurs with previous findings that naively grouping tasks for joint training brings negative transfer among the tasks, thus leading to the incorrect conclusion that MTL does not work.

Task Group	**STL**		**MTL**	
	acc.	f1	acc.	f1
BoolQ	0.806 (±0.128)	0.751 (±0.255)	0.854 (±0.003)	0.843 (±0.004)
RTE	0.829 (±0.011)	0.828 (±0.010)	0.842 (±0.015)	0.841 (±0.014)
CB	0.811 (±0.073)	0.650 (±0.192)	0.925 (±0.039)	0.907 (±0.056)
Causal Language	0.862 (±0.008)	0.825 (±0.010)	0.862 (±0.008)	0.825 (±0.008)
COPA	0.580 (±0.076)	0.552 (±0.124)	0.657 (±0.049)	0.654 (±0.049)
WiC	0.695 (±0.012)	0.689 (±0.015)	0.686 (±0.017)	0.679 (±0.021)
CoLA	0.844 (±0.020)	0.800 (±0.029)	0.846 (±0.020)	0.802 (±0.027)
COPA	0.580 (±0.076)	0.552 (±0.124)	0.678 (±0.139)	0.641 (±0.224)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.961 (±0.002)	0.961 (±0.002)
CB	0.811 (±0.073)	0.650 (±0.192)	0.932 (±0.063)	0.903 (±0.100)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.937 (±0.005)	0.930 (±0.005)
Causal Language	0.862 (±0.008)	0.825 (±0.010)	0.863 (±0.018)	0.827 (±0.020)
CB	0.811 (±0.073)	0.650 (±0.192)	0.904 (±0.040)	0.886 (±0.054)
CoLA	0.844 (±0.020)	0.800 (±0.029)	0.812 (±0.085)	0.719 (±0.217)
COPA	0.580 (±0.076)	0.552 (±0.124)	0.836 (±0.012)	0.790 (±0.018)

Table 3: A comparison of performance between STL and MTL by task groupings with similar PVI estimates for the general and biomedical tasks on the test dataset. Both STL and MTL were fine-tuned using roberta-large.

5.3 Performance of LLMs on the tasks

LLMs are intended to provide solutions for many tasks, thus may be considered multi-task learners. Therefore, we compare our approach to the performance of LLMs as a baseline. Table 7 shows the performance of Llama2-7B, Llama2-70B, and GPT-4 on the general and biomedical tasks using few-shot learning. Overall, our best STL and MTL models performed better than the Llama 2 models for all tasks. On the other hand, GPT-4 shows mixed results in terms of performance compared to the best STL and MTL. It outperformed the best MTL results on RTE, CB, COPA, CoLA and SST2 but underperformed on BoolQ and WiC. It is unclear to what extent these tasks are included in the training corpus of GPT-4. As for the two biomedical

Task Group	STL		MTL	
	acc.	f1	acc.	f1
THYMEMod	0.957 (±0.001)	0.736 (±0.009)	0.958 (±0.001)	0.743 (±0.007)
SeedNeg	0.983 (±0.002)	0.944 (±0.008)	0.986 (±0.002)	0.955 (±0.005)
THYMENeg	0.983 (±0.001)	0.954 (±0.003)	0.984 (±0.001)	0.956 (±0.002)
SeedNeg	0.983 (±0.002)	0.944 (±0.008)	0.986 (±0.004)	0.955 (±0.012)
SeedUncert	0.973 (±0.004)	0.829 (±0.031)	0.977 (±0.006)	0.859 (±0.034)
StratUncert	0.982 (±0.002)	0.675 (±0.071)	0.987 (±0.002)	0.845 (±0.027)
StratNeg	0.987 (±0.002)	0.925 (±0.012)	0.989 (±0.001)	0.941 (±0.004)
SeedNeg	0.983 (±0.002)	0.944 (±0.008)	0.984 (±0.002)	0.949 (±0.005)
THYMEMod	0.957 (±0.001)	0.736 (±0.009)	0.958 (±0.001)	0.736 (±0.012)
SeedUncert	0.973 (±0.004)	0.829 (±0.031)	0.979 (±0.005)	0.868 (±0.029)
StratNeg	0.987 (±0.006)	0.925 (±0.009)	0.985 (±0.003)	0.915 (±0.015)
THYMEMod	0.957 (±0.001)	0.736 (±0.009)	0.958 (±0.002)	0.741 (±0.012)
SeedNeg	0.983 (±0.002)	0.944 (±0.008)	0.983 (±0.002)	0.944 (±0.008)
StratNeg	0.987 (±0.002)	0.925 (±0.012)	0.987 (±0.001)	0.929 (±0.005)

Table 4: A comparison of performance between STL and MTL by task groupings with similar PVI estimates for the clinical tasks on the test dataset. Both STL and MTL were fine-tuned using Bio+Clinical BERT.

Task Group	STL		MTL	
	acc.	f1	acc.	f1
BoolQ	0.806 (±0.128)	0.751 (±0.255)	0.787 (±0.115)	0.729 (±0.240)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.936 (±0.004)	0.929 (±0.005)
CB	0.811 (±0.073)	0.650 (±0.192)	0.846 (±0.046)	0.664 (±0.110)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.936 (±0.006)	0.926 (±0.008)
RTE	0.829 (±0.011)	0.828 (±0.010)	0.734 (±0.134)	0.714 (±0.182)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.938 (±0.010)	0.931 (±0.013)
COPA	0.580 (±0.076)	0.552 (±0.124)	0.514 (±0.046)	0.453 (±0.105)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.938 (±0.005)	0.933 (±0.007)
WiC	0.695 (±0.012)	0.689 (±0.015)	0.674 (±0.026)	0.667 (±0.029)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.939 (±0.003)	0.933 (±0.003)
CoLA	0.844 (±0.020)	0.800 (±0.029)	0.844 (±0.006)	0.798 (±0.008)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.939 (±0.003)	0.932 (±0.006)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.953 (±0.007)	0.953 (±0.007)
Health Advice	0.939 (±0.005)	0.934 (±0.006)	0.929 (±0.029)	0.917 (±0.045)
Causal Language	0.862 (±0.008)	0.825 (±0.010)	0.863 (±0.012)	0.825 (±0.015)
BoolQ	0.806 (±0.128)	0.751 (±0.255)	0.826 (±0.005)	0.812 (±0.006)

Table 5: Performance of MTL by task groupings with different PVI estimates for the general and biomedical datasets. Both STL and MTL were fine-tuned using roberta-large.

tasks, our best MTL model outperformed GPT-4 by a wide margin (0.234-0.331 F1 points), with much higher performance with MTL with task groupings with a similar PVI-based difficulty given the roberta-large model.

Table 8 shows the results of Llama2-7B and Llama2-70B models on the clinical tasks. Similar to the biomedical datasets, the STL and MTL models of the clinical tasks outperformed the Llama 2 models, which indicates that for domain-specific NLP tasks, fine-tuning tasks-specific transformer-based models still works better than prompting the LLMs. Moreover, EMR text is highly unlikely to have been included in the training data of any model given the HIPAA provisions, thus it is ideal for independent evaluations of LLMs.

6 Discussion and Conclusion

In this study, we propose a new method of identifying best-performing task groupings for MTL based on PVI estimates. We conducted experiments on 15 NLP datasets in the general, biomedical, and clinical domains. We showed that when grouping tasks with similar (i.e. not significantly different) PVI estimates, MTL yielded competitive or better results on the majority of the general domain tasks and all biomedical and clinical tasks included in our study. The performance was much better compared to the task grouping with PVI estimates of datasets that are significantly different, suggesting the importance of considering task difficulty as task relatedness and the feasibility of utilizing PVI as a metric for selecting task combinations.

Though MTL may lead to only minor improvements for some tasks as compared to STL, it is more efficient than STL, given the reduced need for training and deployment of multiple models within the computational environment of the average in-

Task Group	STL		MTL	
	acc.	f1	acc.	f1
THYMEMod	0.957 (±0.001)	0.736 (±0.009)	0.956 (±0.002)	0.730 (±0.008)
StratUncert	0.982 (±0.002)	0.675 (±0.071)	0.984 (±0.003)	0.813 (±0.030)
THYMENeg	0.983 (±0.001)	0.954 (±0.003)	0.984 (±0.001)	0.957 (±0.003)
THYMEMod	0.957 (±0.001)	0.736 (±0.009)	0.958 (±0.001)	0.740 (±0.008)
SeedNeg	0.938 (±0.002)	0.944 (±0.008)	0.979 (±0.003)	0.936 (±0.008)
StratUncert	0.982 (±0.002)	0.675 (±0.071)	0.982 (±0.004)	0.671 (±0.108)
SeedUncert	0.973 (±0.004)	0.829 (±0.031)	0.972 (±0.008)	0.795 (±0.068)
StratNeg	0.987 (±0.002)	0.925 (±0.012)	0.985 (±0.002)	0.919 (±0.013)
StratNeg	0.987 (±0.002)	0.925 (±0.012)	0.986 (±0.001)	0.919 (±0.007)
THYMENeg	0.983 (±0.001)	0.954 (±0.003)	0.983 (±0.001)	0.953 (±0.002)
StratUncert	0.982 (±0.002)	0.675 (±0.071)	0.981 (±0.001)	0.636 (±0.033)
THYMENeg	0.983 (±0.001)	0.954 (±0.003)	0.983 (±0.001)	0.953 (±0.002)

Table 6: Performance of MTL by task groupings with different PVI estimates for the clinical datasets. Both STL and MTL were fine-tuned using Bio+Clinical BERT.

stitution, e.g. medical academic center or hospital. For instance, applying two separate roberta-large models for the two biomedical tasks in our experiment would require tuning around 710 million parameters (355 million for each model). In contrast, MTL with hard parameter sharing involves tuning the shared 355 million parameters once, plus a small percentage of task-specific parameters, with the total number of parameters significantly less than 710 million. While LLMs yielded varying results on the general domain datasets, both STL and MTL models consistently outperformed Llama 2 and GPT-4 models in the biomedical domain. This indicates that for domain-specific tasks, fine-tuned models may remain a preferable option. Due to the computational capacity needed for fine-tuning LLMs and the data privacy regulation when using the LLMs, the PVI-based method of grouping tasks for MTL could be particularly beneficial for domain-specific tasks.

This study focuses on task grouping for MTL using PVI. This metric could also be applied for data instance selection to optimize training sets for joint learning. Moreover, investigating parameter sharing and monitoring the variation in weights for both single learners and joint learning models, by ranking parameters based on their change after grouping tasks, represents another avenue for improved MTL's effectiveness and efficiency. We will explore these directions in future work.

We believe our proposed approach could be incorporated into frameworks for stacking LMs to algorithmically optimize LM prompting, fine-tuning, augmentation, and reasoning, e.g. DSPy framework[4] (Khattab et al., 2022, 2023).

Dataset	Model	Acc.	F1
BoolQ	Llama2-7B	0.507	0.309
	Llama2-70B	0.625	0.378
	GPT-4	0.697	0.591
CB	Llama2-7B	0.250	0.258
	Llama2-70B	0.446	0.323
	GPT-4	0.964	0.932
RTE	Llama2-7B	0.498	0.332
	Llama2-70B	0.783	0.522
	GPT-4	0.895	0.894
COPA	Llama2-7B	0.450	0.227
	Llama2-70B	0.495	0.331
	GPT-4	0.940	0.940
WiC	Llama2-7B	0.517	0.337
	Llama2-70B	0.586	0.576
	GPT-4	0.502	0.358
CoLA	Llama2-7B	0.249	0.178
	Llama2-70B	0.740	0.646
	GPT-4	0.856	0.829
SST2	Llama2-7B	0.232	0.179
	Llama2-70B	0.930	0.930
	GPT-4	0.964	0.964
Health Advice	Llama2-7B	0.316	0.213
	Llama2-70B	0.205	0.185
	GPT-4	0.719	0.700
Causal Langauge	Llama2-7B	0.165	0.113
	Llama2-70B	0.282	0.255
	GPT-4	0.545	0.494

Table 7: Performance of Llama2-7B, Llama2-70B, and GPT-4 on the general and biomedical tasks.

Dataset	Model	Acc.	F1
THYMENeg	Llama2-7B	0.309	0.196
	Llama2-70B	0.379	0.342
THYMEMod	Llama2-7B	0.673	0.184
	Llama2-70B	0.177	0.073
SeedNeg	Llama2-7B	0.462	0.246
	Llama2-70B	0.618	0.428
SeedUncert	Llama2-7B	0.738	0.299
	Llama2-70B	0.204	0.126
StratNeg	Llama2-7B	0.226	0.139
	Llama2-70B	0.487	0.369
StratUncert	Llama2-7B	0.397	0.204
	Llama2-70B	0.690	0.420

Table 8: Performance of Llama2-7B and Llama2-70B on the clinical tasks.

[4]https://github.com/stanfordnlp/dspy

Limitations

As noted, one limitation of the current study is that we were not able to test GPT-4 performance on the clinical tasks because we do not have access to a HIPAA-compliant GPT-4 model and are not allowed to transmit data to the GPT4 public API.

We chose to experiment with two of the most competitive LLMs - the Llama 2 family of models which are locally downloadable and GPT-4. Experiments with other LLMs can be pursued, however we believe the LLMs results we report in this study are likely representative of the general trends.

The results of the single learners are with smaller models (less than 7B parameters) as fine-tuning and PVI estimates are computationally feasible given our computational resources. We encourage those with more abundant computational resources to experiment with our methodology using LLMs as the base models which we believe will further improve our reported results.

Ethics Statement

The clinical datasets we worked with represent a Limited Data Set where all confidential data are removed except for dates. We did not transmit any part of these clinical datasets to any public API and processed the dataset locally on a HIPAA-compliant server. We have an approved IRB for the study described in this paper.

Acknowledgements

The work is funded by grant R01GM114355 from the US National Institutes of Health.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Armen Aghajanyan, Anchit Gupta, Akshat Shrivastava, Xilun Chen, Luke Zettlemoyer, and Sonal Gupta. 2021. Muppet: Massive multi-task representations with pre-finetuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5799–5811.

Hector Martinez Alonso and Barbara Plank. 2017. When is multitask learning effective? semantic sequence prediction under varying data conditions. In *EACL 2017-15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–10.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q Tran, Dara Bahri, Jianmo Ni, et al. 2021. Ext5: Towards extreme multi-task scaling for transfer learning. *arXiv preprint arXiv:2111.10952*.

Joachim Bingel and Anders Søgaard. 2017. Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 164–169.

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

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28:41–75.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of NAACL-HLT*, pages 2924–2936.

Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. 2023. Learning sample difficulty from pre-trained models for reliable prediction. *arXiv preprint arXiv:2304.10127*.

Marie-Catherine De Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. In *proceedings of Sinn und Bedeutung*, volume 23, pages 107–124.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. 2022. Understanding dataset difficulty with $\mathcal{V}$-usable information. In

International Conference on Machine Learning, pages 5988–6008. PMLR.

Chris Fifty, Ehsan Amid, Zhe Zhao, Tianhe Yu, Rohan Anil, and Chelsea Finn. 2021. Efficiently identifying task groupings for multi-task learning. *Advances in Neural Information Processing Systems*, 34:27503–27516.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. 2020. Learning to branch for multi-task learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3854–3863.

Michael Hahn, Dan Jurafsky, and Richard Futrell. 2021. Sensitivity as a complexity measure for sequence classification tasks. *Transactions of the Association for Computational Linguistics*, 9:891–908.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Aditya Khandelwal and Benita Kathleen Britto. 2020. Multitask learning of negation and speculation using transformers. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 79–87, Online. Association for Computational Linguistics.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Artur Kulmizev and Joakim Nivre. 2023. Investigating ud treebanks via dataset difficulty measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1076–1089.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Dongyue Li, Huy Nguyen, and Hongyang Ryan Zhang. 2023a. Identification of negative transfers in multitask learning using surrogate models. *Transactions on Machine Learning Research*.

Dongyue Li, Jinhong Yu, and Hongyang Zhang. 2023b. Approximate clustering for extracting task relationships in multi-instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Yingya Li, Jun Wang, and Bei Yu. 2021. Detecting health advice in medical research literature. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6018–6029, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. 2023. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1463–1476, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sheng Lu, Shan Chen, Yingya Li, Danielle Bitterman, Guergana Savova, and Iryna Gurevych. 2023. Measuring pointwise $\mathcal{V}$-usable information in-context-ly. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15739–15756, Singapore. Association for Computational Linguistics.

Jingwei Ni, Zhijing Jin, Qian Wang, Mrinmaya Sachan, and Markus Leippold. 2023. When does aggregating multiple skills with multi-task learning work? a case study in financial NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7465–7488, Toronto, Canada. Association for Computational Linguistics.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. An empirical study of multi-task learning on BERT for biomedical text mining. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 205–214, Online. Association for Computational Linguistics.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. Rissanen data analysis: Examining dataset characteristics via description length. In *International Conference on Machine Learning*, pages 8500–8513. PMLR.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. *arXiv preprint arXiv:2304.10703*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. 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.

Xiaozhuang Song, Shun Zheng, Wei Cao, James Yu, and Jiang Bian. 2022. Efficient and effective multi-task grouping via meta learning on task combinations. *Advances in Neural Information Processing Systems*, 35:37647–37659.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *International conference on machine learning*, pages 9120–9132. PMLR.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Antonio Torralba and Alexei A. Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528.

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

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

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

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114.

Sen Wu, Hongyang R Zhang, and Christopher Ré. 2019. Understanding and improving information transfer in multi-task learning. In *International Conference on Learning Representations*.

Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. 2020. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*.

Bei Yu, Yingya Li, and Jun Wang. 2019. Detecting causal language use in science findings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4664–4674, Hong Kong, China. Association for Computational Linguistics.

Chen Zhang, Qiuchi Li, Dawei Song, and Benyou Wang. 2020a. A multi-task learning framework for opinion triplet extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 819–828, Online. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Haau-Sing Li, and Samuel R Bowman. 2020b. When do you need billions of words of pretraining data? *arXiv preprint arXiv:2011.04946*.

Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.

Xinran Zhao, Shikhar Murty, and Christopher Manning. 2022. On measuring the intrinsic few-shot hardness of datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3955–3963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Appendix

A.1 Experiment details

We used an NVIDIA Titan RTX GPU cluster of 7 nodes for the single- and multi-task training experiments. We experimented with training epochs of {5, 6, 7, 8, 9, 10}, random seeds of {42, 52, 62, 72, 82}, a max sequence length of {126, 512}, a learning rate of {1e-5, 2e-5}, with a fixed batch size of 32 and gradient accumulation steps of 2. All experiments with LLMs were run using an NVIDIA A100 or through OpenAI API. Table 9 provides details of the datasets we used. Table 10 shows the label distribution of each dataset.

dataset	domain	train | dev | test sets size	type of task	number of classes
BoolQ	General	7541 | 1886 | 3270	question and answering	2
CB	General	200 | 50 | 57	natural language inference	3
RTE	General	1992 | 498 | 277	text entailment	2
COPA	General	320 | 80 | 100	question and answering	2
WiC	General	4342 | 1086 | 638	word sentence disambiguation	2
CoLA	General	8640 | 1711 | 1043	acceptability	2
SST2	General	53879 | 13470 | 872	sentiment analysis	2
Health Advice	Biomedical	3828 | 957 | 1197	suggestion mining	3
Causal Language	Biomedical	1958 | 490 | 614	causal relation	4
THYMENeg	Clinical	28637 | 7160 | 19240	negation detection	2
THYMEMod	Clinical	28637 | 7160 | 19240	contextual modality detection	4
SeedNeg	Clinical	4666 | 1167 | 676	negation detection	2
SeedUncert	Clinical	4666 | 1167 | 676	uncertainty detection	2
StratNeg	Clinical	3568 | 802 | 1737	negation detection	2
StratUncert	Clinical	3568 | 802 | 1737	uncertainty detection	2

Table 9: Overview of datasets used in this study. The datasets cover a variety of tasks and domains, providing a comprehensive base for our analyses.

task	label	train	dev	test
BoolQ	True	4699	1175	2033
	False	2842	711	1237
CB	contradiction	95	24	28
	entailment	92	23	23
	neutral	13	3	5
RTE	entailment	999	250	146
	not_entailment	993	248	131
COPA	0	160	40	50
	1	160	40	50
WiC	True	2171	543	319
	False	2171	543	319
CoLA	0	2022	506	322
	1	4818	1205	721
SST2	0	23824	5956	428
	1	30055	7514	444
Health Advice	no_advice	2296	574	705
	weak_advice	943	236	303
	strong_advice	589	147	189
Causal Language	no_relation	860	215	281
	correlation	639	160	199
	conditional_causation	320	80	94
	direct_causation	139	35	39
THYMENeg	0	3079	770	1961
	1	25558	6390	17279
THYMEMod	acutal	26358	6590	17750
	hypothetical	1236	309	781
	hedged	656	164	417
	generic	387	97	292
SeedNeg	0	3960	991	619
	1	704	176	57
SeedUncert	0	4466	1117	647
	1	198	50	29
StratNeg	0	3033	759	1658
	1	173	43	80
StratUncert	0	3118	780	1700
	1	88	22	38

Table 10: Label distribution of each dataset.

Task Group	**Ours**		**TaskEmb**	
	acc.	f1	acc.	f1
BoolQ	0.854 (±0.003)	0.843 (±0.004)	0.830 (±0.005)	0.817 (±0.005)
CB	0.925 (±0.039)	0.907 (±0.056)	0.925 (±0.053)	0.899 (±0.090)
RTE	0.842 (±0.015)	0.841 (±0.014)	0.588 (±0.147)	0.520 (±0.237)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.775 (±0.302)	0.706 (±0.419)
COPA	0.836 (±0.012)	0.790 (±0.018)	0.505 (±0.014)	0.365 (±0.090)
WiC	0.686 (±0.017)	0.679 (±0.021)	0.540 (±0.076)	0.433 (±0.175)
CoLA	0.846 (±0.020)	0.802 (±0.027)	0.802 (±0.083)	0.700 (±0.209)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.862 (±0.245)	0.827 (±0.341)
Health Advice	0.937 (±0.005)	0.930 (±0.005)	0.937 (±0.005)	0.930 (±0.005)
Causal Language	0.863 (±0.018)	0.827 (±0.020)	0.863 (±0.018)	0.827 (±0.020)

Table 11: Performance of our approach and baseline-TaskEmbed on the general and biomedical datasets. Both STL and MTL were fine-tuned using roberta-large.

Task Group	Ours		TaskEmb	
	acc.	f1	acc.	f1
THYMEMod	0.958 (±0.001)	0.743 (±0.007)	0.958 (±0.001)	0.740 (±0.008)
THYMENeg	0.984 (±0.001)	0.956 (±0.002)	0.984 (±0.001)	0.957 (±0.003)
SeedNeg	0.986 (±0.002)	0.955 (±0.005)	0.982 (±0.002)	0.943 (±0.006)
SeedUncert	0.979 (±0.005)	0.868 (±0.029)	0.977 (±0.003)	0.855 (±0.031)
StratNeg	0.989 (±0.001)	0.941 (±0.004)	0.984 (±0.002)	0.915 (±0.011)
StratUncert	0.987 (±0.002)	0.845 (±0.027)	0.985 (±0.002)	0.792 (±0.037)

Table 12: Performance of our approach and baseline-TaskEmbed on the clinical datasets. Both STL and MTL were fine-tuned using Bio+Clinical BERT.

Task Group	Ours		Surrogate Models	
	acc.	f1	acc.	f1
BoolQ	0.854 (±0.003)	0.843 (±0.004)	0.830 (±0.005)	0.817 (±0.005)
CB	0.925 (±0.039)	0.907 (±0.056)	0.925 (±0.053)	0.899 (±0.090)
CB	0.925 (±0.039)	0.907 (±0.056)	0.932 (±0.063)	0.903 (±0.100)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.961 (±0.002)	0.961 (±0.002)
RTE	0.842 (±0.015)	0.841 (±0.014)	0.657 (±0.180)	0.590 (±0.291)
COPA	0.836 (±0.012)	0.790 (±0.018)	0.578 (±0.105)	0.500 (±0.197)
COPA	0.836 (±0.012)	0.790 (±0.018)	0.611 (±0.126)	0.544 (±0.239)
BoolQ	0.854 (±0.003)	0.843 (±0.004)	0.699 (±0.131)	0.551 (±0.285)
WiC	0.686 (±0.017)	0.679 (±0.021)	0.540 (±0.076)	0.433 (±0.175)
COPA	0.836 (±0.012)	0.790 (±0.018)	0.505 (±0.014)	0.365 (±0.090)
CoLA	0.846 (±0.020)	0.802 (±0.027)	0.787 (±0.108)	0.647 (±0.270)
CB	0.925 (±0.039)	0.907 (±0.056)	0.793 (±0.183)	0.694 (±0.290)
SST2	0.961 (±0.002)	0.961 (±0.002)	0.867 (±0.249)	0.833 (±0.344)
WiC	0.686 (±0.017)	0.679 (±0.021)	0.625 (±0.087)	0.587 (±0.177)
Health Advice	0.937 (±0.005)	0.930 (±0.005)	0.937 (±0.005)	0.930 (±0.005)
Causal Language	0.863 (±0.018)	0.827 (±0.020)	0.863 (±0.018)	0.827 (±0.020)

Table 13: Performance of our approach and baseline-Surrogate Models on the general and biomedical datasets. Both STL and MTL were fine-tuned using roberta-large.

Task Group	Ours		Surrogate Models	
	acc.	f1	acc.	f1
THYMEMod	0.958 (±0.001)	0.743 (±0.007)	0.958 (±0.001)	0.740 (±0.008)
THYMENeg	0.984 (±0.001)	0.956 (±0.002)	0.984 (±0.001)	0.957 (±0.003)
SeedNeg	0.986 (±0.002)	0.955 (±0.005)	0.983 (±0.004)	0.944 (±0.012)
THYMENeg	0.984 (±0.001)	0.956 (±0.002)	0.984 (±0.001)	0.955 (±0.002)
SeedUncert	0.979 (±0.005)	0.868 (±0.029)	0.977 (±0.006)	0.859 (±0.034)
StratUncert	0.987 (±0.002)	0.845 (±0.027)	0.987 (±0.002)	0.845 (±0.027)
StratNeg	0.989 (±0.001)	0.941 (±0.004)	0.984 (±0.002)	0.915 (±0.011)
StratUncert	0.987 (±0.002)	0.845 (±0.027)	0.985 (±0.002)	0.792 (±0.037)

Table 14: Performance of our approach and baseline-Surrogate Models on the clinical datasets. Both STL and MTL were fine-tuned using Bio+Clinical BERT.

Task Group	STL		MTL	
	acc.	f1	acc.	f1
BoolQ	0.688 (±0.003)	0.663 (±0.005)	0.704 (±0.014)	0.682 (±0.015)
RTE	0.612 (±0.026)	0.606 (±0.026)	0.632 (±0.040)	0.626 (±0.045)
CB	0.682 (±0.009)	0.477 (±0.007)	0.861 (±0.019)	0.836 (±0.070)
WiC	0.619 (±0.010)	0.607 (±0.012)	0.627 (±0.007)	0.615 (±0.009)
CoLA	0.815 (±0.006)	0.754 (±0.008)	0.820 (±0.003)	0.765 (±0.007)
COPA	0.526 (±0.047)	0.524 (±0.047)	0.539 (±0.023)	0.534 (±0.026))
SST2	0.922 (±0.005)	0.922 (±0.005)	0.921 (±0.008)	0.921 (±0.008)
CB	0.682 (±0.009)	0.477 (±0.007)	0.839 (±0.057)	0.760 (±0.091)
Health Advice	0.928 (±0.004)	0.917 (±0.005)	0.930 (±0.005)	0.922 (±0.006)
Causal Language	0.823 (±0.016)	0.750 (±0.066)	0.860 (±0.008)	0.829 (±0.007)

Table 15: A comparison of performance between STL and MTL by task groupings with similar PVI estimates for the general and biomedical tasks on the test dataset with average result and a 95% confidence interval of 5 random seed values samples, where the t value for 95% confidence is 2.776. Both STL and MTL were fine-tuned using bert-base-uncased.

Task Group	**STL**		**MTL**	
	acc.	f1	acc.	f1
THYMENeg	0.983 (±0.001)	0.953 (±0.002)	0.983 (±0.001)	0.954 (±0.002)
SeedNeg	0.960 (±0.010)	0.875 (±0.024)	0.970 (±0.009)	0.905 (±0.027)
THYMEMod	0.959 (±0.001)	0.761 (±0.009)	0.960 (±0.0001)	0.760 (±0.010)
SeedNeg	0.960 (±0.010)	0.875 (±0.024)	0.961 (±0.009)	0.881 (±0.022)
SeedNeg	0.960 (±0.010)	0.875 (±0.024)	0.962 (±0.008)	0.881 (±0.022)
StratNeg	0.983 (±0.004)	0.905 (±0.020)	0.987 (±0.001)	0.939 (±0.009)
SeedUncert	0.973 (±0.004)	0.809 (±0.025)	0.974 (±0.003)	0.823 (±0.014)
StratUncert	0.980 (±0.002)	0.589 (±0.085)	0.985 (±0.003)	0.775 (±0.043)

Table 16: A comparison of performance between STL and MTL by task groupings with similar PVI estimates for the clinical tasks on the test dataset with average result and 95% confidence interval of 5 random seed values samples, where the t value for a 95% confidence is 2.776. Both STL and MTL were fine-tuned using BioBERT.

A.2 Example of prompts

Tables 17 and 18 are examples of the prompts used in the current study for Llama 2 and GPT-4.

SENTENCE1: Then the silence in the Zoo became complete. Woil stared around him and then suddenly with a push of his wings raised himself into the air, turned, and landed ten feet away on the back of a green bench. Creggan could see that he was afraid and that his fear was making him terribly uncertain.
SENTENCE2: Woil was afraid
QUESTION: Is this (0) entailment, (1) contradiction, or (2) neutral?
ANSWER: 0

SENTENCE1: He's weird enough to have undressed me without thinking, according to some mad notion of the "proper" thing to do. Perhaps he thought I couldn't lie in bed with my clothes on.
SENTENCE2: she couldn't lie in bed with her clothes on
QUESTION: Is this (0) entailment, (1) contradiction, or (2) neutral?
ANSWER: 1

SENTENCE1: I hope you are settling down and the cat is well. This was a lie. She did not hope the cat was well.
SENTENCE2: the cat was well
QUESTION: Is this (0) entailment, (1) contradiction, or (2) neutral?
ANSWER: 2

CONTEXT: Dietary interventions in older people were effective in maintaining fruit and fish intake, but this did not lead to a significant reduction in cognitive decline.
QUESTION: Is this (0) no relationship, (1) correlation, (2) conditional causation, or (3) direct causation?
ANSWER: 3

Table 17: An example of a 4-shot prompt for the CB task.

CONTEXT: Further research is needed to evaluate clinical applications of the PA, such as a more accurate identification of malnourished cardiac surgery patients.
QUESTION: Is this (0) no relationship, (1) correlation, (2) conditional causation, or (3) direct causation?
ANSWER: 0

CONTEXT: The etiology of anemia appears to be iron-related and precipitated by the female sex.
QUESTION: Is this (0) no relationship, (1) correlation, (2) conditional causation, or (3) direct causation?
ANSWER: 1

CONTEXT: Diet may influence the pharmacokinetics of ASA, but effects may be through modulation of glycine conjugation rather than glucuronidation.
QUESTION: Is this (0) no relationship, (1) correlation, (2) conditional causation, or (3) direct causation?
ANSWER: 2

CONTEXT: Dietary interventions in older people were effective in maintaining fruit and fish intake, but this did not lead to a significant reduction in cognitive decline.
QUESTION: Is this (0) no relationship, (1) correlation, (2) conditional causation, or (3) direct causation?
ANSWER: 3

Table 18: An example of a 4-shot prompt for the Causal Language task.