On the Potential of Multi-Task Learning in Predictive Process Monitoring -Supplementary Material-

Lukas Kirchdorfer[1,2], **Keyvan Amiri Elyasi**[2], **Heiner Stuckenschmidt**[2]

[1]SAP Signavio, Walldorf, Germany
[2]Data and Web Science Group, University of Mannheim, Germany
lukas.kirchdorfer@@sap.com, {keyvan.amiri, heiner.stuckenschmidt}@uni-mannheim.de

Introduction

This document provides comprehensive supplementary material accompanying our main paper. It details the implementation settings, experimental design, and additional empirical results that support and extend the findings reported in the main text. We include exhaustive information to ensure full transparency and reproducibility of our work.

The document is structured into two main parts. The first part, *Implementation Details*, outlines our experimental setup, including hyperparameter tuning, computing infrastructure, and a short description of employed datasets. We elaborate on the grid search procedures conducted for learning rate and specific hyperparameters for some of the multi-task optimization (MTO) methods to provide information relevant for reproducing the experiments.

The second part, *Additional Results*, presents a broad range of empirical findings and statistical analyses. We begin by reporting statistical tests comparing the sensitivity of different model architectures to multi-task learning (MTL), followed by tests that assess the relative performance of various MTO methods. Furthermore, we include the complete experimental results for all dataset-architecture-task combinations in tabular format.

To complement our main findings, we further investigate the role of task combinations in MTL and provide visualizations that illustrate their effects on model performance for both LSTM and Transformer architectures. We also report the results of an ablation study conducted on a simplified CNN model. Additionally, we delve into two specific aspects of next activity prediction: (i) the performance gains for infrequent classes, and (ii) the impact of prefix length. These analyses highlight where and how MTL provides significant benefits in real-world process prediction scenarios.

To facilitate navigation, we provide a structured index below. Each section is self-contained and can be consulted independently, depending on the reader's focus.

Contents

1 Implementation Details

In this section, we explain the applied settings used for the reported experiments in more detail.

Experimental Setup

Hyperparameters. As described in the main paper, we perform an extensive grid search over the learning rate and MTO-specific hyperparameters. Here, we provide more details on the search spaces for the latter.

For UW-SO, we tune the softmax temperature parameter $T \in \{1, 3, 5, 7, 10, 15, 20\}$, controlling the uniformity of task weights. In the case of Scalarization, we evaluate all valid combinations of task weights $\omega_k \in \{0.1, 0.2, \ldots, 0.9\}$ such that $\sum_k \omega_k = 1$, yielding 9 combinations for two tasks and 36 for three tasks. For GradNorm, we follow Xin et al. (2022) and search over $\alpha \in \{0.5, 1.0, 1.5, 2.0\}$, controlling the training rate balance. Lastly, for CAGrad, we adopt a slightly reduced hyperparameter grid for the convergence rate compared to Liu et al. (2021), using $c \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$.

Furthermore, we use three random seeds, which are $42, 123, 2025$.

We report the best hyperparameters for the three-task combination and the CNN model in Tables 1 and 2.

Computing Infrastructure. Regarding the computing infrastructure, all experiments were conducted on a Linux machine using an Nvidia RTX A6000 GPU with 48GB VRAM. Our Python code (version 3.9) is based on PyTorch 2.7.0.

Datasets. In the following, we provide additional details on the datasets used in our study. The selected datasets capture

Table 1: Best learning rates and MTO-specific hyperparameters for three-task combination with CNN (Part 1).

Dataset	MTO Method	Learning Rate	MTO HPO
BPIC13I	UW-SO	0.0001	1.0
BPIC13I	RLW	0.0001	/
BPIC13I	CAGrad	0.0001	0.5
BPIC13I	EW	0.0001	/
BPIC13I	GradDrop	0.0001	/
BPIC13I	DWA	0.0001	/
BPIC13I	UW-O	0.0001	/
BPIC13I	NashMTL	0.0001	/
BPIC13I	GradNorm	0.0001	1.5
BPIC13I	PCGrad	0.0001	/
BPIC13I	GLS	0.0001	/
BPIC13I	UW	0.0001	/
BPIC13I	IMTL	0.0001	/
P2P	Scalarization	0.001	[0.2, 0.2, 0.6]
P2P	UW-SO	0.001	15.0
P2P	DWA	0.001	/
P2P	RLW	0.001	/
P2P	GradDrop	0.001	/
P2P	CAGrad	0.001	0.1
P2P	GradNorm	0.001	1.5
P2P	NashMTL	0.0001	/
P2P	IMTL	0.001	/
P2P	UW-O	0.001	/
P2P	EW	0.001	/
P2P	GLS	0.001	/
P2P	PCGrad	0.001	/
P2P	UW	0.001	/
Production	UW-SO	0.001	15.0
Production	DWA	0.001	/
Production	RLW	0.001	/
Production	GradDrop	0.001	/
Production	CAGrad	0.001	0.1
Production	GradNorm	0.001	1.5
Production	NashMTL	0.001	/
Production	IMTL	0.001	/
Production	UW-O	0.001	/
Production	EW	0.001	/
Production	GLS	0.001	/
Production	PCGrad	0.001	/
Production	UW	0.001	/
HelpDesk	UW-SO	0.001	3.0
HelpDesk	DWA	0.001	/
HelpDesk	RLW	0.001	/
HelpDesk	GradDrop	0.001	/
HelpDesk	CAGrad	0.001	0.5
HelpDesk	NashMTL	0.001	/
HelpDesk	IMTL	0.001	/
HelpDesk	UW-O	0.001	/
HelpDesk	EW	0.001	/
HelpDesk	PCGrad	0.001	/
HelpDesk	UW	0.001	/
HelpDesk	GradNorm	0.001	2.0
HelpDesk	GLS	0.001	/
Sepsis	UW-SO	0.001	3.0
Sepsis	DWA	0.001	/
Sepsis	RLW	0.001	/
Sepsis	GradDrop	0.001	/
Sepsis	CAGrad	0.001	0.3
Sepsis	GradNorm	0.0001	2.0

Table 2: Best learning rates and MTO-specific hyperparameters for three-task combination with CNN (Part 2).

Dataset	MTO Method	Learning Rate	MTO HPO
Sepsis	NashMTL	0.0001	/
Sepsis	IMTL	0.001	/
Sepsis	EW	0.001	/
Sepsis	GLS	0.001	/
Sepsis	PCGrad	0.001	/
Sepsis	UW	0.001	/
Sepsis	UW-O	0.001	/
BPIC20DD	UW-SO	0.001	10.0
BPIC20DD	DWA	0.001	/
BPIC20DD	RLW	0.001	/
BPIC20DD	GradDrop	0.001	/
BPIC20DD	CAGrad	0.001	0.1
BPIC20DD	GradNorm	0.001	2.0
BPIC20DD	NashMTL	0.001	/
BPIC20DD	IMTL	0.001	/
BPIC20DD	EW	0.001	/
BPIC20DD	GLS	0.001	/
BPIC20DD	PCGrad	0.001	/
BPIC20DD	UW	0.001	/
BPIC20DD	UW-O	0.001	/
BPIC15-1	UW-SO	0.0001	20.0
BPIC15-1	DWA	0.0001	/
BPIC15-1	RLW	0.0001	/
BPIC15-1	GradDrop	0.0001	/
BPIC15-1	CAGrad	0.0001	0.1
BPIC15-1	GradNorm	0.0001	2.0
BPIC15-1	NashMTL	0.0001	/
BPIC15-1	IMTL	0.0001	/
BPIC15-1	EW	0.0001	/
BPIC15-1	PCGrad	0.0001	/
BPIC15-1	UW	0.0001	/
BPIC15-1	GLS	0.0001	/
BPIC15-1	UW-O	0.0001	/
BPIC20ID	UW-SO	0.0001	10.0
BPIC20ID	DWA	0.0001	/
BPIC20ID	RLW	0.0001	/
BPIC20ID	GradDrop	0.0001	/
BPIC20ID	CAGrad	0.0001	0.1
BPIC20ID	GradNorm	0.0001	2.0
BPIC20ID	NashMTL	0.0001	/
BPIC20ID	IMTL	0.0001	/
BPIC20ID	EW	0.0001	/
BPIC20ID	PCGrad	0.0001	/
BPIC20ID	UW	0.0001	/
BPIC20ID	UW-O	0.0001	/
BPIC20ID	GLS	0.0001	/
BPI12C	UW-SO	0.0001	20.0
BPI12C	RLW	0.0001	/
BPI12C	GradDrop	0.0001	/
BPI12C	CAGrad	0.0001	0.1
BPI12C	DWA	0.0001	/
BPI12C	IMTL	0.0001	/
BPI12C	EW	0.0001	/
BPI12C	UW	0.0001	/
BPI12C	GradNorm	0.0001	1.5
BPI12C	PCGrad	0.0001	/
BPI12C	UW-O	0.0001	/
BPI12C	GLS	0.0001	/
BPI12C	NashMTL	0.0001	/

process execution data from a diverse range of domains, including manufacturing, healthcare, IT services, finance and public services. This diversity enables a more comprehensive assessment of the impact of MTL, as we go beyond the limited scope of earlier studies that often rely on just two or three datasets or focus on a single domain. All datasets are publicly available[1] and have been included in our code appendix.

In the following, we provide a short description for each dataset:

- BPIC20DD: The dataset contains two years of travel expense claim events from a university—2017 covering two departments, and 2018 covering the entire institution. The process involves submitting domestic declarations, which follow the following workflow: employee submission, approval by travel administration, budget owner, and supervisor (with steps skipped if roles overlap), and occasionally the director. The process concludes with either a trip or a payment. Domestic trips require no prior approval. Claims for reimbursement can be filed either before or after the trip, depending on the expense type.
- BPIC20ID: The same process as BPIC20DD, but this one contains international instead of domestic declarations. International trips need supervisor-approved permits.
- BPIC15-1: This dataset, provided by a Dutch municipality, contains four years of building permit applications. It includes a wide range of activities labeled with both codes and task names in Dutch and English. Cases cover the main application process and potential objection procedures. Additional information includes the responsible resource and the application cost.
- BPIC13I: Log of Volvo IT incident management system.
- BPIC12C: This dataset was provided by a Dutch financial institute, and it contains data from a loan application process.
- Helpdesk: Event log concerning the ticketing management process of the Help desk of an Italian software company
- Sepsis: This real-life event log records the hospital pathways of sepsis patients, based on data from the hospital's ERP system. Sepsis is a life-threatening condition typically caused by an infection. Each event is enriched with 39 attributes, such as responsible groups, test results, and checklist information. All data is anonymized, and while event timestamps are randomized, the relative timing within each case is preserved.
- P2P: This event log captures a purchase-to-pay process within an organization. Each case represents a procurement transaction. The process includes steps such as creating and analyzing requests for quotation, comparing supplier offers, setting conditions, and creating purchase orders. Approval, confirmation, delivery, invoicing, and payment activities are also recorded.
- Production: The Production event log captures detailed manufacturing process executions over time. Each case represents the production of a specific item or order and consists of a sequence of activities (e.g., Turning & Milling, Laser Marking, Grinding, Inspection, Packing).

[1] https://data.4tu.nl/datasets

2 Additional Results

Statistical tests for comparing model architectures

In this section, we provide the results of statistical tests used to compare how much different architectures are impacted by MTL.

We used the Wilcoxon signed-rank test on the Δ_m values. To do so, we aggregated the results over all 13 MTO methods and all task combinations (i.e., NAP+NTP+RTP, NAP+NTP, NAP+RTP, NTP+RTP). The results show a statistically significant difference between LSTM and CNN ($p = 0.0391$), suggesting that these architectures respond differently to MTL. However, no significant differences were found between LSTM and Transformer ($p = 0.5469$) or CNN and Transformer ($p = 0.25$).

Statistical tests for comparing MTO methods

In this section, we provide the results of statistical tests that are used to compare different MTO methods.

We used the Friedman test to compare different MTO methods. To do so, we have aggregated the results over all model architectures (i.e., LSTM, CNN, and Transformer) and all task combinations (i.e., NAP+NTP+RTP, NAP+NTP, NAP+RTP, NTP+RTP). The Friedman test indicated significant differences across MTO methods ($\chi^2 = 26.42$, $p = 0.0093$). However, the post-hoc Nemenyi test did not reveal any significant pairwise differences at the $p < 0.05$ level (see Table 3). Still, Wilcoxon signed-rank tests comparing each method to the baseline 'EW' showed that 'CAGrad' significantly outperformed 'EW' ($p = 0.0391$), while no other statistically significant difference is observed between other MTO methods and equal weighting.

Main results

To complement the main paper, we provide all experimental results obtained in this study in the following. For each dataset, we report a separate table per combination of architecture and tasks.

Table 3: P-values from the post-hoc Nemenyi test comparing MTO strategies following a significant Friedman test.

	EW	DWA	RLW	UW	UW-SO	UW-O	GLS	GradDrop	CAGrad	PCGrad	GradNorm	IMTL	Nash-MTL
EW	1.0000	1.0000	0.9996	1.0000	0.5564	1.0000	1.0000	1.0000	0.7398	0.9727	0.9813	1.0000	0.9998
DWA	1.0000	1.0000	0.9951	0.9999	0.7398	1.0000	1.0000	0.9992	0.8813	0.9951	0.9971	0.9998	1.0000
RLW	0.9996	0.9951	1.0000	1.0000	0.0824	0.9614	0.9999	1.0000	0.1634	0.5087	0.5564	1.0000	0.8813
UW	1.0000	0.9999	1.0000	1.0000	0.2210	0.9971	1.0000	1.0000	0.3719	0.7804	0.8178	1.0000	0.9813
UW-SO	0.5564	0.7398	0.0824	0.2210	1.0000	0.9071	0.4617	0.1390	1.0000	0.9998	0.9996	0.1907	0.9727
UW-O	1.0000	1.0000	0.9614	0.9971	0.9071	1.0000	1.0000	0.9876	0.9727	0.9998	0.9999	0.9951	1.0000
GLS	1.0000	1.0000	0.9999	1.0000	0.4617	1.0000	1.0000	1.0000	0.6510	0.9469	0.9614	1.0000	0.9992
GradDrop	1.0000	0.9992	1.0000	1.0000	0.1390	0.9876	1.0000	1.0000	0.2544	0.6510	0.6964	1.0000	0.9469
CAGrad	0.7398	0.8813	0.1634	0.3719	1.0000	0.9727	0.6510	0.2544	1.0000	1.0000	1.0000	0.3301	0.9951
PCGrad	0.9727	0.9951	0.5087	0.7804	0.9998	0.9998	0.9469	0.6510	1.0000	1.0000	1.0000	0.7398	1.0000
GradNorm	0.9813	0.9971	0.5564	0.8178	0.9996	0.9999	0.9614	0.6964	1.0000	1.0000	1.0000	0.7804	1.0000
IMTL	1.0000	0.9998	1.0000	1.0000	0.1907	0.9951	1.0000	1.0000	0.3301	0.7398	0.7804	1.0000	0.9727
Nash-MTL	0.9998	1.0000	0.8813	0.9813	0.9727	1.0000	0.9992	0.9469	0.9951	1.0000	1.0000	0.9727	1.0000

Table 4: Results on the P2P dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.73 (0.01)	1.07 (0.08)	4.47 (0.11)	/	/
EW	0.86 (0.00)	1.08 (0.07)	4.73 (0.15)	46	-4.19
Scalarization	0.86 (0.00)	1.03 (0.06)	4.47 (0.20)	76	-7.17
DWA	0.86 (0.00)	1.06 (0.07)	4.60 (0.38)	56	-5.46
RLW	0.86 (0.01)	1.06 (0.08)	4.73 (0.51)	64	-4.64
UW	0.87 (0.01)	1.08 (0.05)	5.02 (0.23)	35	-2.07
UW-O	0.86 (0.01)	1.05 (0.07)	5.05 (0.21)	42	-2.48
UW-SO	0.87 (0.01)	1.06 (0.08)	4.50 (0.29)	58	-6.56
GLS	0.86 (0.01)	1.09 (0.06)	5.05 (0.18)	44	-1.52
IMTL	0.85 (0.01)	1.07 (0.08)	5.13 (0.52)	41	-1.09
GradNorm	0.87 (0.01)	1.06 (0.04)	4.53 (0.05)	75	-6.17
NashMTL	0.86 (0.00)	1.04 (0.07)	4.48 (0.48)	63	-7.06
GradDrop	0.87 (0.01)	1.03 (0.02)	4.41 (0.09)	69	-8.05
PCGrad	0.87 (0.01)	1.03 (0.02)	4.47 (0.33)	74	-7.66
CAGrad	0.86 (0.01)	1.08 (0.07)	4.49 (0.18)	61	-5.90

Table 5: Results on the P2P dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.73 (0.01)	1.07 (0.08)	/	/
EW	0.87 (0.00)	1.08 (0.08)	14	-8.99
Scalarization	0.86 (0.01)	1.09 (0.07)	19	-8.55
DWA	0.87 (0.01)	1.06 (0.09)	15	-9.94
RLW	0.86 (0.01)	1.06 (0.07)	19	-9.84
UW	0.86 (0.02)	1.07 (0.07)	27	-8.92
UW-O	0.86 (0.02)	1.08 (0.09)	16	-8.99
UW-SO	0.87 (0.01)	1.08 (0.09)	21	-9.32
GLS	0.86 (0.02)	1.08 (0.08)	20	-8.94
IMTL	0.86 (0.02)	1.09 (0.09)	16	-8.53
GradNorm	0.87 (0.01)	1.08 (0.08)	16	-9.19
NashMTL	0.87 (0.01)	1.07 (0.07)	15	-9.59
GradDrop	0.86 (0.02)	1.07 (0.06)	16	-8.89
PCGrad	0.86 (0.02)	1.06 (0.07)	16	-9.71
CAGrad	0.86 (0.02)	1.06 (0.08)	16	-9.49

Table 6: Results on the P2P dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.73 (0.01)	4.47 (0.11)	/	/
EW	0.86 (0.00)	4.46 (0.25)	69	-9.26
Scalarization	0.86 (0.00)	4.49 (0.40)	69	-8.78
DWA	0.86 (0.01)	4.59 (0.08)	60	-7.54
RLW	0.86 (0.00)	4.47 (0.21)	76	-9.18
UW	0.86 (0.02)	4.99 (0.27)	33	-3.04
UW-O	0.85 (0.01)	5.00 (0.23)	48	-2.44
UW-SO	0.86 (0.01)	4.40 (0.19)	69	-9.91
GLS	0.85 (0.01)	5.02 (0.24)	45	-2.53
IMTL	0.85 (0.01)	4.86 (0.38)	54	-4.26
GradNorm	0.86 (0.01)	4.49 (0.14)	64	-8.49
NashMTL	0.87 (0.01)	4.53 (0.18)	62	-8.72
GradDrop	0.86 (0.00)	4.45 (0.12)	73	-9.48
PCGrad	0.87 (0.02)	4.67 (0.19)	48	-7.29
CAGrad	0.86 (0.00)	4.68 (0.27)	61	-6.44

Table 7: Results on the P2P dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.07 (0.08)	4.47 (0.11)	/	/
EW	1.07 (0.02)	4.38 (0.38)	65	-1.00
DWA	1.06 (0.05)	4.39 (0.15)	61	-1.53
RLW	1.08 (0.03)	4.61 (0.22)	57	1.77
UW	1.09 (0.04)	4.56 (0.23)	60	1.66
UW-O	1.06 (0.05)	4.58 (0.50)	68	0.66
UW-SO	1.04 (0.06)	4.34 (0.24)	106	-3.06
GLS	1.09 (0.05)	4.47 (0.21)	82	0.58
IMTL	1.06 (0.05)	4.39 (0.25)	69	-1.81
GradNorm	1.06 (0.05)	4.45 (0.29)	83	-0.99
NashMTL	1.04 (0.08)	4.30 (0.30)	72	-3.40
GradDrop	1.05 (0.06)	4.47 (0.43)	64	-1.07
PCGrad	1.04 (0.03)	4.33 (0.11)	80	-3.28
CAGrad	1.05 (0.02)	4.44 (0.25)	76	-1.32

Table 8: Results on the P2P dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.29 (0.00)	1.13 (0.11)	5.40 (0.20)	/	/
EW	0.86 (0.01)	1.10 (0.05)	5.47 (0.14)	55	-65.08
Scalarization	0.86 (0.00)	1.14 (0.06)	5.44 (0.29)	56	-64.25
DWA	0.86 (0.01)	1.11 (0.08)	5.50 (0.19)	35	-64.75
RLW	0.86 (0.01)	1.11 (0.08)	5.60 (0.21)	45	-64.17
UW	0.86 (0.01)	1.12 (0.07)	5.67 (0.18)	26	-63.34
UW-O	0.86 (0.01)	1.12 (0.07)	5.98 (0.04)	29	-60.79
UW-SO	0.86 (0.01)	1.11 (0.08)	5.57 (0.19)	44	-64.23
GLS	0.86 (0.02)	1.15 (0.08)	5.88 (0.12)	35	-60.86
IMTL	0.86 (0.00)	1.12 (0.10)	5.89 (0.26)	28	-61.63
GradNorm	0.87 (0.01)	1.10 (0.09)	5.61 (0.08)	40	-65.07
NashMTL	0.85 (0.01)	1.14 (0.07)	5.70 (0.12)	92	-60.79
GradDrop	0.86 (0.02)	1.13 (0.06)	5.66 (0.23)	29	-62.54
PCGrad	0.86 (0.01)	1.12 (0.09)	5.62 (0.33)	30	-63.56
CAGrad	0.86 (0.01)	1.11 (0.07)	5.63 (0.14)	29	-63.73

Table 9: Results on the P2P dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.13 (0.11)	5.40 (0.20)	/	/
EW	1.13 (0.09)	5.46 (0.25)	45	0.65
DWA	1.16 (0.07)	5.55 (0.24)	28	2.93
RLW	1.15 (0.10)	5.62 (0.18)	33	3.01
UW	1.16 (0.10)	5.62 (0.20)	42	3.35
UW-O	1.10 (0.07)	5.72 (0.14)	34	1.77
UW-SO	1.13 (0.06)	5.49 (0.13)	41	0.84
GLS	1.11 (0.09)	5.70 (0.18)	34	1.71
IMTL	1.11 (0.08)	5.67 (0.20)	28	1.66
GradNorm	1.14 (0.08)	5.59 (0.12)	40	2.24
NashMTL	1.19 (0.10)	5.70 (0.16)	88	5.61
GradDrop	1.15 (0.07)	5.60 (0.12)	31	2.60
PCGrad	1.13 (0.07)	5.53 (0.11)	38	1.30
CAGrad	1.13 (0.09)	5.45 (0.30)	34	0.43

Table 10: Results on the P2P dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.29 (0.00)	5.40 (0.20)	/	/
EW	0.86 (0.01)	5.57 (0.08)	41	-94.84
Scalarization	0.85 (0.01)	5.42 (0.15)	63	-94.65
DWA	0.86 (0.00)	5.43 (0.15)	47	-95.80
RLW	0.86 (0.01)	5.54 (0.13)	43	-96.34
UW	0.85 (0.01)	5.79 (0.26)	27	-92.16
UW-O	0.86 (0.01)	6.07 (0.16)	26	-90.03
UW-SO	0.86 (0.01)	5.61 (0.18)	44	-94.86
GLS	0.85 (0.00)	6.10 (0.12)	24	-88.85
IMTL	0.86 (0.01)	5.86 (0.23)	22	-92.60
GradNorm	0.85 (0.00)	5.50 (0.01)	34	-94.41
NashMTL	0.84 (0.01)	5.72 (0.20)	87	-91.11
GradDrop	0.86 (0.02)	5.56 (0.23)	35	-95.58
PCGrad	0.86 (0.02)	5.51 (0.12)	48	-96.47
CAGrad	0.86 (0.01)	5.65 (0.28)	26	-95.43

Table 11: Results on the P2P dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.29 (0.00)	1.13 (0.11)	/	/
EW	0.86 (0.02)	1.13 (0.08)	17	-96.16
Scalarization	0.85 (0.01)	1.14 (0.08)	15	-95.21
DWA	0.85 (0.01)	1.13 (0.08)	18	-95.45
RLW	0.85 (0.02)	1.13 (0.09)	23	-94.87
UW	0.85 (0.01)	1.13 (0.07)	56	-95.99
UW-O	0.86 (0.01)	1.15 (0.07)	8	-95.59
UW-SO	0.86 (0.01)	1.12 (0.07)	63	-96.73
GLS	0.85 (0.01)	1.15 (0.06)	13	-94.65
IMTL	0.85 (0.01)	1.12 (0.07)	60	-96.16
GradNorm	0.86 (0.01)	1.12 (0.07)	24	-97.67
NashMTL	0.85 (0.01)	1.12 (0.08)	14	-95.67
GradDrop	0.86 (0.00)	1.12 (0.08)	14	-96.84
PCGrad	0.85 (0.01)	1.12 (0.06)	59	-96.35
CAGrad	0.85 (0.02)	1.13 (0.08)	14	-95.71

Table 12: Results on the P2P dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.58 (0.03)	1.12 (0.08)	5.47 (0.28)	/	/
EW	0.86 (0.00)	1.13 (0.06)	5.32 (0.12)	37	-17.15
DWA	0.86 (0.01)	1.12 (0.07)	5.40 (0.05)	30	-16.93
RLW	0.87 (0.01)	1.12 (0.06)	5.54 (0.22)	45	-16.55
UW	0.87 (0.01)	1.08 (0.09)	5.58 (0.29)	49	-17.47
UW-O	0.87 (0.01)	1.10 (0.04)	5.92 (0.28)	39	-14.66
UW-SO	0.86 (0.00)	1.14 (0.06)	5.45 (0.11)	29	-16.07
GLS	0.86 (0.01)	1.18 (0.04)	5.67 (0.28)	40	-13.76
IMTL	0.86 (0.00)	1.13 (0.05)	6.22 (0.48)	30	-11.76
GradNorm	0.86 (0.01)	1.13 (0.06)	5.57 (0.05)	32	-15.71
NashMTL	0.86 (0.00)	1.11 (0.06)	5.20 (0.18)	38	-18.37
GradDrop	0.84 (0.01)	1.19 (0.09)	5.42 (0.11)	32	-13.34
PCGrad	0.86 (0.01)	1.09 (0.06)	5.63 (0.10)	45	-16.29
CAGrad	0.87 (0.01)	1.12 (0.05)	5.45 (0.12)	31	-17.15

Table 13: Results on the P2P dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.58 (0.03)	1.12 (0.08)	/	/
EW	0.86 (0.02)	1.10 (0.07)	24	-25.17
Scalarization	0.86 (0.01)	1.09 (0.07)	26	-26.30
DWA	0.86 (0.02)	1.11 (0.05)	24	-25.44
RLW	0.86 (0.01)	1.10 (0.08)	26	-25.32
UW	0.87 (0.01)	1.12 (0.07)	18	-25.25
UW-O	0.86 (0.01)	1.12 (0.04)	35	-24.56
UW-SO	0.86 (0.01)	1.11 (0.06)	27	-25.08
GLS	0.85 (0.01)	1.11 (0.08)	24	-24.79
IMTL	0.85 (0.02)	1.11 (0.05)	15	-24.43
GradNorm	0.87 (0.00)	1.08 (0.08)	30	-27.16
NashMTL	0.86 (0.01)	1.08 (0.09)	24	-26.33
GradDrop	0.87 (0.01)	1.10 (0.08)	20	-26.24
PCGrad	0.87 (0.00)	1.09 (0.06)	23	-27.00
CAGrad	0.87 (0.00)	1.08 (0.07)	21	-27.43

Table 14: Results on the P2P dataset for Transformer with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.58 (0.03)	5.47 (0.28)	/	/
EW	0.85 (0.01)	5.52 (0.11)	24	-23.48
Scalarization	0.85 (0.02)	5.20 (0.28)	55	-26.01
DWA	0.86 (0.02)	5.38 (0.12)	34	-25.09
RLW	0.86 (0.01)	5.49 (0.20)	35	-24.24
UW	0.86 (0.01)	5.45 (0.04)	36	-24.79
UW-O	0.86 (0.01)	5.59 (0.05)	41	-23.38
UW-SO	0.86 (0.02)	5.42 (0.24)	38	-24.83
GLS	0.86 (0.01)	5.71 (0.06)	36	-22.71
IMTL	0.86 (0.02)	5.64 (0.22)	47	-23.09
GradNorm	0.86 (0.01)	5.45 (0.35)	39	-24.47
NashMTL	0.85 (0.01)	5.38 (0.25)	28	-24.82
GradDrop	0.82 (0.04)	5.20 (0.36)	44	-23.71
PCGrad	0.86 (0.01)	5.48 (0.28)	31	-24.46
CAGrad	0.86 (0.02)	5.53 (0.28)	30	-24.44

Table 15: Results on the P2P dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.12 (0.08)	5.47 (0.28)	/	/
EW	1.19 (0.08)	5.19 (0.10)	44	0.51
DWA	1.19 (0.04)	5.06 (0.35)	39	-0.86
RLW	1.21 (0.00)	5.35 (0.18)	29	3.10
UW	1.21 (0.07)	5.27 (0.08)	31	2.17
UW-O	1.15 (0.05)	5.44 (0.11)	30	0.76
UW-SO	1.16 (0.03)	5.17 (0.27)	34	-0.99
GLS	1.14 (0.09)	5.51 (0.21)	20	1.19
IMTL	1.18 (0.10)	5.40 (0.32)	42	2.09
GradNorm	1.16 (0.10)	5.25 (0.15)	51	-0.15
NashMTL	1.22 (0.03)	5.55 (0.38)	22	5.11
GradDrop	1.27 (0.11)	5.21 (0.22)	37	4.39
PCGrad	1.17 (0.06)	5.31 (0.11)	29	0.66
CAGrad	1.16 (0.08)	5.30 (0.23)	34	0.24

Table 16: Results on the Production dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.44 (0.01)	0.99 (0.16)	4.72 (0.29)	/	/
EW	0.49 (0.02)	1.01 (0.15)	4.77 (0.15)	64	-2.50
DWA	0.49 (0.01)	1.02 (0.15)	4.77 (0.39)	70	-2.19
RLW	0.49 (0.03)	1.02 (0.14)	4.79 (0.28)	79	-2.27
UW	0.49 (0.02)	1.03 (0.14)	5.05 (0.26)	63	0.15
UW-O	0.49 (0.02)	1.02 (0.15)	5.63 (0.38)	48	3.94
UW-SO	0.51 (0.02)	1.01 (0.16)	4.85 (0.07)	65	-3.08
GLS	0.50 (0.02)	1.04 (0.16)	5.37 (0.13)	58	2.17
IMTL	0.50 (0.02)	1.08 (0.14)	6.69 (2.10)	51	12.89
GradNorm	0.49 (0.01)	1.06 (0.19)	6.09 (0.03)	50	9.08
NashMTL	0.50 (0.02)	1.02 (0.17)	4.85 (0.12)	64	-2.50
GradDrop	0.50 (0.02)	1.00 (0.15)	4.79 (0.03)	71	-3.44
PCGrad	0.50 (0.02)	1.01 (0.15)	4.87 (0.17)	61	-2.17
CAGrad	0.49 (0.02)	1.08 (0.17)	4.98 (0.41)	66	1.50

Table 17: Results on the Production dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.44 (0.01)	0.99 (0.16)	/	/
EW	0.52 (0.01)	0.99 (0.16)	16	-8.41
Scalarization	0.51 (0.02)	1.00 (0.16)	18	-6.95
DWA	0.51 (0.02)	0.99 (0.16)	17	-7.83
RLW	0.52 (0.01)	0.99 (0.16)	18	-7.89
UW	0.52 (0.01)	0.99 (0.16)	18	-8.62
UW-O	0.51 (0.01)	0.99 (0.16)	16	-7.48
UW-SO	0.52 (0.03)	0.99 (0.16)	17	-8.12
GLS	0.51 (0.03)	0.99 (0.16)	17	-7.98
IMTL	0.52 (0.01)	1.00 (0.16)	22	-7.59
GradNorm	0.52 (0.01)	0.99 (0.16)	20	-8.16
NashMTL	0.51 (0.02)	0.99 (0.16)	17	-7.13
GradDrop	0.50 (0.02)	0.99 (0.16)	17	-6.95
PCGrad	0.52 (0.03)	0.99 (0.16)	17	-8.29
CAGrad	0.51 (0.02)	0.99 (0.16)	17	-7.80

Table 18: Results on the Production dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.44 (0.01)	4.72 (0.29)	/	/
EW	0.49 (0.01)	4.92 (0.23)	68	-2.99
Scalarization	0.49 (0.02)	4.59 (0.28)	81	-6.37
DWA	0.51 (0.02)	5.07 (0.39)	54	-3.58
RLW	0.49 (0.02)	4.86 (0.15)	77	-3.38
UW	0.50 (0.03)	5.16 (0.27)	62	-1.81
UW-O	0.48 (0.02)	5.47 (0.31)	56	3.39
UW-SO	0.50 (0.02)	4.85 (0.13)	62	-5.50
GLS	0.49 (0.01)	5.54 (0.09)	55	2.86
IMTL	0.48 (0.03)	5.65 (0.20)	62	5.48
GradNorm	0.48 (0.02)	5.02 (0.31)	80	-1.24
NashMTL	0.48 (0.02)	4.65 (0.26)	80	-5.56
GradDrop	0.48 (0.02)	4.62 (0.36)	79	-5.84
PCGrad	0.49 (0.03)	4.93 (0.34)	67	-3.25
CAGrad	0.49 (0.01)	5.10 (0.24)	53	-1.67

Table 19: Results on the Production dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	0.99 (0.16)	4.72 (0.29)	/	/
EW	1.02 (0.16)	4.71 (0.13)	69	1.67
Scalarization	1.01 (0.14)	4.52 (0.19)	83	-1.10
DWA	1.03 (0.16)	4.86 (0.20)	68	3.49
RLW	1.01 (0.14)	4.89 (0.24)	69	2.82
UW	1.06 (0.16)	4.91 (0.19)	70	5.94
UW-O	1.06 (0.15)	4.53 (0.20)	95	1.66
UW-SO	1.06 (0.12)	4.76 (0.26)	70	4.20
GLS	1.07 (0.15)	4.85 (0.30)	71	5.74
IMTL	1.12 (0.12)	4.57 (0.22)	94	5.13
GradNorm	0.99 (0.16)	6.19 (0.36)	150	16.00
NashMTL	1.03 (0.19)	4.73 (0.44)	73	2.44
GradDrop	1.00 (0.16)	4.76 (0.12)	59	1.01
PCGrad	1.05 (0.14)	4.76 (0.50)	79	3.42
CAGrad	1.11 (0.14)	4.59 (0.43)	118	4.92

Table 20: Results on the Production dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.29 (0.00)	0.99 (0.16)	10.15 (1.36)	/	/
EW	0.44 (0.02)	1.02 (0.16)	7.15 (0.64)	133	-25.04
DWA	0.42 (0.04)	1.04 (0.19)	9.84 (4.94)	85	-13.75
RLW	0.45 (0.01)	1.02 (0.17)	6.77 (0.13)	114	-27.50
UW	0.44 (0.02)	1.01 (0.16)	8.10 (1.20)	105	-22.75
UW-O	0.43 (0.02)	1.02 (0.17)	9.28 (0.06)	34	-17.79
UW-SO	0.43 (0.02)	1.01 (0.16)	8.15 (1.44)	101	-21.48
GLS	0.44 (0.01)	1.02 (0.15)	9.62 (0.40)	43	-17.66
IMTL	0.43 (0.00)	1.09 (0.16)	9.31 (0.60)	43	-14.36
GradNorm	0.43 (0.01)	1.02 (0.14)	10.35 (1.23)	67	-13.41
NashMTL	0.44 (0.02)	1.01 (0.16)	7.50 (0.36)	52	-25.07
GradDrop	0.44 (0.03)	1.00 (0.16)	8.75 (2.11)	109	-21.14
PCGrad	0.43 (0.02)	1.03 (0.14)	8.51 (1.75)	87	-19.92
CAGrad	0.44 (0.00)	1.03 (0.20)	8.42 (0.79)	111	-20.38

Table 21: Results on the Production dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	0.99 (0.16)	10.15 (1.36)	/	/
EW	1.06 (0.06)	10.22 (6.06)	70	3.64
Scalarization	1.02 (0.16)	7.32 (0.75)	114	-12.51
DWA	1.00 (0.15)	9.22 (3.38)	61	-4.22
RLW	1.01 (0.18)	12.33 (3.17)	58	11.75
UW	1.00 (0.16)	10.72 (1.05)	76	2.89
UW-O	1.04 (0.18)	6.68 (0.47)	82	-15.00
UW-SO	1.01 (0.16)	8.86 (4.23)	76	-5.47
GLS	1.06 (0.19)	6.82 (0.22)	105	-13.15
IMTL	1.09 (0.23)	8.92 (1.53)	57	-1.13
GradNorm	1.01 (0.17)	9.06 (0.42)	92	-4.72
NashMTL	1.02 (0.16)	6.76 (0.31)	86	-15.47
GradDrop	1.00 (0.17)	8.97 (3.31)	64	-5.55
PCGrad	1.04 (0.11)	10.37 (6.00)	49	3.47
CAGrad	1.08 (0.17)	7.40 (0.46)	107	-9.05

Table 22: Results on the Production dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.29 (0.00)	10.15 (1.36)	/	/
EW	0.44 (0.01)	8.60 (3.03)	102	-31.94
Scalarization	0.44 (0.01)	6.98 (0.78)	152	-39.97
DWA	0.44 (0.01)	8.58 (2.81)	95	-32.01
RLW	0.44 (0.03)	8.52 (1.03)	135	-31.95
UW	0.42 (0.02)	10.59 (3.11)	76	-19.36
UW-O	0.45 (0.02)	9.91 (0.07)	28	-27.24
UW-SO	0.45 (0.01)	7.03 (0.19)	140	-42.29
GLS	0.45 (0.01)	10.79 (0.27)	93	-23.38
IMTL	0.43 (0.01)	11.03 (0.63)	26	-19.30
GradNorm	0.43 (0.01)	9.78 (3.00)	96	-24.71
NashMTL	0.44 (0.01)	7.58 (0.51)	56	-37.50
GradDrop	0.40 (0.08)	9.80 (3.70)	85	-20.22
PCGrad	0.40 (0.07)	9.91 (5.21)	79	-19.13
CAGrad	0.43 (0.01)	10.50 (0.83)	81	-20.84

Table 23: Results on the Production dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.29 (0.00)	0.99 (0.16)	/	/
EW	0.42 (0.01)	1.00 (0.16)	12	-21.26
Scalarization	0.44 (0.01)	1.01 (0.16)	13	-23.45
DWA	0.43 (0.02)	1.01 (0.16)	43	-22.18
RLW	0.43 (0.00)	1.02 (0.16)	13	-22.25
UW	0.43 (0.01)	1.00 (0.16)	19	-23.13
UW-O	0.44 (0.01)	1.02 (0.14)	11	-24.19
UW-SO	0.43 (0.01)	1.00 (0.16)	16	-22.49
GLS	0.44 (0.00)	1.02 (0.15)	12	-23.41
IMTL	0.44 (0.02)	1.01 (0.15)	10	-24.01
GradNorm	0.44 (0.00)	1.01 (0.16)	12	-23.95
NashMTL	0.43 (0.01)	1.02 (0.16)	7	-21.39
GradDrop	0.43 (0.02)	1.00 (0.16)	19	-23.31
PCGrad	0.42 (0.00)	1.00 (0.16)	16	-21.93
CAGrad	0.43 (0.00)	1.01 (0.16)	11	-22.75

Table 24: Results on the Production dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.38 (0.02)	0.99 (0.16)	6.80 (1.99)	/	/
EW	0.51 (0.03)	0.99 (0.16)	6.40 (1.42)	41	-12.65
Scalarization	0.51 (0.01)	0.99 (0.16)	5.40 (0.65)	60	-17.73
DWA	0.49 (0.01)	0.99 (0.16)	7.13 (0.31)	23	-7.98
RLW	0.51 (0.02)	0.99 (0.16)	6.06 (0.47)	55	-14.36
UW	0.51 (0.02)	0.99 (0.16)	6.24 (1.16)	41	-13.53
UW-O	0.51 (0.01)	1.00 (0.16)	6.22 (0.32)	54	-13.87
UW-SO	0.52 (0.02)	0.99 (0.16)	5.73 (0.28)	51	-16.90
GLS	0.54 (0.00)	1.00 (0.16)	6.49 (0.77)	49	-15.13
IMTL	0.51 (0.03)	1.06 (0.13)	6.76 (0.95)	38	-8.85
GradNorm	0.52 (0.01)	0.99 (0.16)	6.28 (0.56)	43	-14.43
NashMTL	0.51 (0.01)	1.00 (0.15)	5.05 (0.68)	50	-19.93
GradDrop	0.49 (0.02)	0.99 (0.16)	6.02 (0.59)	50	-13.16
PCGrad	0.53 (0.01)	0.99 (0.16)	5.66 (0.46)	53	-18.00
CAGrad	0.53 (0.02)	1.00 (0.16)	5.96 (1.02)	80	-16.28

Table 25: Results on the Production dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.38 (0.02)	0.99 (0.16)	/	/
EW	0.55 (0.01)	0.99 (0.16)	14	-21.57
Scalarization	0.55 (0.01)	1.00 (0.15)	16	-21.41
DWA	0.55 (0.01)	0.99 (0.16)	14	-21.71
RLW	0.55 (0.01)	0.99 (0.15)	13	-22.10
UW	0.55 (0.01)	0.99 (0.16)	14	-21.42
UW-O	0.55 (0.01)	0.99 (0.16)	17	-21.65
UW-SO	0.54 (0.00)	1.00 (0.15)	14	-21.15
GLS	0.54 (0.01)	0.99 (0.15)	15	-20.53
IMTL	0.54 (0.02)	1.00 (0.15)	18	-20.65
GradNorm	0.55 (0.01)	0.99 (0.16)	14	-21.47
NashMTL	0.55 (0.00)	0.99 (0.16)	15	-21.81
GradDrop	0.55 (0.01)	0.99 (0.16)	14	-21.63
PCGrad	0.55 (0.01)	1.00 (0.16)	14	-21.50
CAGrad	0.55 (0.01)	0.99 (0.16)	14	-21.63

Table 26: Results on the Production dataset for Transformer with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.38 (0.02)	6.80 (1.99)	/	/
EW	0.50 (0.01)	7.55 (0.91)	25	-9.68
Scalarization	0.44 (0.01)	5.95 (0.40)	44	-13.32
DWA	0.51 (0.01)	6.86 (0.59)	36	-16.05
RLW	0.50 (0.02)	6.74 (1.08)	38	-15.77
UW	0.51 (0.01)	6.56 (0.94)	39	-18.88
UW-O	0.52 (0.01)	5.89 (0.22)	50	-24.52
UW-SO	0.51 (0.01)	6.28 (0.89)	42	-20.20
GLS	0.53 (0.00)	7.34 (1.48)	41	-16.10
IMTL	0.49 (0.02)	6.77 (0.43)	39	-15.09
GradNorm	0.52 (0.01)	6.33 (0.30)	42	-21.47
NashMTL	0.50 (0.01)	5.28 (0.62)	42	-26.18
GradDrop	0.51 (0.00)	5.64 (0.12)	58	-24.85
PCGrad	0.51 (0.02)	7.34 (0.82)	26	-13.11
CAGrad	0.51 (0.01)	6.49 (0.57)	57	-19.10

Table 27: Results on the Production dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	0.99 (0.16)	6.80 (1.99)	/	/
EW	0.99 (0.16)	5.76 (0.67)	54	-7.44
Scalarization	0.99 (0.16)	5.21 (0.45)	60	-11.56
DWA	0.99 (0.16)	5.84 (0.83)	55	-6.91
RLW	0.99 (0.16)	5.65 (0.46)	54	-8.25
UW	0.99 (0.16)	6.11 (0.16)	44	-4.88
UW-O	1.00 (0.16)	5.64 (0.73)	61	-8.02
UW-SO	0.99 (0.16)	5.84 (0.79)	54	-6.83
GLS	1.00 (0.16)	5.85 (0.18)	58	-6.40
IMTL	0.96 (0.22)	6.63 (0.56)	35	-2.71
GradNorm	0.99 (0.16)	5.43 (0.62)	57	-10.01
NashMTL	0.99 (0.16)	5.51 (1.18)	52	-9.22
GradDrop	0.99 (0.16)	5.70 (0.52)	54	-7.94
PCGrad	0.99 (0.16)	5.90 (1.22)	58	-6.50
CAGrad	1.00 (0.15)	6.13 (1.06)	45	-4.43

Table 28: Results on the Helpdesk dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.68 (0.01)	5.09 (0.10)	6.38 (0.06)	/	/
EW	0.80 (0.00)	5.07 (0.15)	6.33 (0.10)	34	-6.43
DWA	0.80 (0.01)	5.09 (0.14)	6.32 (0.10)	38	-6.39
RLW	0.80 (0.01)	5.08 (0.09)	6.33 (0.06)	43	-6.30
UW	0.80 (0.01)	5.10 (0.13)	6.34 (0.08)	39	-6.20
UW-O	0.80 (0.00)	5.07 (0.12)	6.37 (0.03)	18	-6.17
UW-SO	0.80 (0.00)	5.06 (0.12)	6.35 (0.12)	37	-6.42
GLS	0.80 (0.00)	5.09 (0.10)	6.35 (0.07)	26	-6.20
IMTL	0.80 (0.01)	5.10 (0.09)	6.35 (0.05)	34	-5.98
GradNorm	0.80 (0.00)	5.06 (0.10)	6.35 (0.04)	43	-6.34
NashMTL	0.80 (0.00)	5.09 (0.12)	6.31 (0.09)	39	-6.39
GradDrop	0.80 (0.01)	5.06 (0.10)	6.33 (0.08)	46	-6.50
PCGrad	0.80 (0.00)	5.07 (0.09)	6.33 (0.06)	27	-6.47
CAGrad	0.80 (0.00)	5.07 (0.14)	6.34 (0.10)	31	-6.26

Table 29: Results on the Helpdesk dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.68 (0.01)	5.09 (0.10)	/	/
EW	0.80 (0.00)	5.06 (0.14)	24	-9.52
Scalarization	0.80 (0.01)	5.05 (0.12)	22	-9.63
DWA	0.80 (0.00)	5.08 (0.12)	31	-9.31
RLW	0.80 (0.00)	5.08 (0.14)	32	-9.26
UW	0.80 (0.01)	5.09 (0.16)	22	-9.27
UW-O	0.80 (0.00)	5.06 (0.13)	19	-9.63
UW-SO	0.80 (0.01)	5.07 (0.15)	22	-9.37
GLS	0.80 (0.00)	5.08 (0.13)	22	-9.30
IMTL	0.80 (0.01)	5.09 (0.11)	22	-8.97
GradNorm	0.80 (0.00)	5.09 (0.08)	23	-9.30
NashMTL	0.80 (0.00)	5.08 (0.11)	33	-9.44
GradDrop	0.80 (0.01)	5.07 (0.16)	24	-9.22
PCGrad	0.80 (0.00)	5.06 (0.14)	22	-9.45
CAGrad	0.80 (0.01)	5.05 (0.10)	25	-9.76

Table 30: Results on the Helpdesk dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.68 (0.01)	6.38 (0.06)	/	/
EW	0.80 (0.00)	6.31 (0.09)	36	-9.54
Scalarization	0.80 (0.01)	6.32 (0.08)	41	-9.48
DWA	0.80 (0.00)	6.32 (0.07)	42	-9.50
RLW	0.80 (0.00)	6.33 (0.06)	46	-9.56
UW	0.79 (0.01)	6.39 (0.08)	25	-8.43
UW-O	0.80 (0.01)	6.37 (0.06)	28	-9.10
UW-SO	0.80 (0.00)	6.34 (0.08)	43	-9.35
GLS	0.80 (0.01)	6.36 (0.06)	21	-9.04
IMTL	0.80 (0.00)	6.40 (0.05)	18	-8.93
GradNorm	0.80 (0.00)	6.35 (0.07)	32	-9.40
NashMTL	0.80 (0.00)	6.37 (0.06)	37	-8.86
GradDrop	0.79 (0.01)	6.35 (0.09)	50	-8.96
PCGrad	0.80 (0.01)	6.33 (0.04)	44	-9.24
CAGrad	0.80 (0.00)	6.38 (0.03)	31	-9.07

Table 31: Results on the Helpdesk dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	5.09 (0.10)	6.38 (0.06)	/	/
EW	5.12 (0.08)	6.37 (0.03)	42	0.20
DWA	5.15 (0.11)	6.37 (0.05)	35	0.55
RLW	5.18 (0.20)	6.49 (0.18)	35	1.75
UW	5.13 (0.09)	6.38 (0.05)	49	0.36
UW-O	5.13 (0.08)	6.38 (0.06)	42	0.41
UW-SO	5.07 (0.12)	6.34 (0.04)	42	-0.46
GLS	5.09 (0.10)	6.34 (0.09)	42	-0.32
IMTL	5.10 (0.10)	6.35 (0.11)	52	-0.14
GradNorm	5.09 (0.08)	6.32 (0.08)	54	-0.44
NashMTL	5.09 (0.09)	6.34 (0.10)	45	-0.31
GradDrop	5.11 (0.13)	6.32 (0.07)	47	-0.29
PCGrad	5.08 (0.11)	6.30 (0.07)	46	-0.72
CAGrad	5.08 (0.10)	6.36 (0.11)	39	-0.28

Table 32: Results on the Helpdesk dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.66 (0.01)	5.58 (0.07)	7.01 (0.11)	/	/
EW	0.77 (0.01)	5.55 (0.08)	6.97 (0.19)	32	-5.81
DWA	0.77 (0.01)	5.48 (0.05)	6.88 (0.15)	49	-6.86
RLW	0.77 (0.00)	5.52 (0.05)	6.95 (0.15)	59	-6.20
UW	0.78 (0.00)	5.42 (0.12)	6.86 (0.14)	57	-7.75
UW-O	0.79 (0.00)	5.40 (0.09)	6.77 (0.06)	45	-8.51
UW-SO	0.78 (0.01)	5.43 (0.09)	6.86 (0.12)	57	-7.47
GLS	0.79 (0.00)	5.40 (0.09)	6.80 (0.13)	41	-8.22
IMTL	0.78 (0.01)	5.55 (0.21)	6.93 (0.17)	37	-6.39
GradNorm	0.78 (0.00)	5.38 (0.13)	6.78 (0.07)	47	-8.24
NashMTL	0.78 (0.00)	5.50 (0.08)	6.92 (0.02)	46	-6.55
GradDrop	0.77 (0.01)	5.58 (0.06)	7.00 (0.12)	46	-5.42
PCGrad	0.78 (0.00)	5.50 (0.17)	6.89 (0.14)	42	-6.75
CAGrad	0.79 (0.01)	5.40 (0.13)	6.79 (0.09)	50	-8.25

Table 33: Results on the Helpdesk dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	5.58 (0.07)	7.01 (0.11)	/	/
EW	5.57 (0.12)	6.93 (0.09)	142	-0.66
DWA	5.53 (0.05)	6.95 (0.08)	41	-0.85
RLW	5.62 (0.17)	7.02 (0.06)	36	0.40
UW	5.57 (0.06)	6.89 (0.07)	54	-0.91
UW-O	5.52 (0.08)	6.91 (0.01)	43	-1.23
UW-SO	5.48 (0.09)	6.90 (0.07)	55	-1.72
GLS	5.55 (0.09)	6.93 (0.09)	153	-0.81
IMTL	5.52 (0.07)	6.98 (0.06)	40	-0.72
GradNorm	5.55 (0.08)	6.92 (0.06)	159	-0.85
NashMTL	5.53 (0.04)	6.98 (0.07)	42	-0.61
GradDrop	5.55 (0.15)	6.94 (0.09)	146	-0.82
PCGrad	5.56 (0.12)	6.95 (0.05)	52	-0.59
CAGrad	5.49 (0.11)	7.00 (0.08)	52	-0.84

Table 34: Results on the Helpdesk dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.66 (0.01)	7.01 (0.11)	/	/
EW	0.78 (0.00)	6.75 (0.08)	59	-10.67
Scalarization	0.78 (0.00)	6.71 (0.06)	48	-11.26
DWA	0.78 (0.00)	6.73 (0.01)	66	-10.83
RLW	0.78 (0.01)	6.94 (0.11)	44	-9.00
UW	0.78 (0.00)	6.86 (0.12)	51	-9.77
UW-O	0.78 (0.00)	6.83 (0.14)	37	-10.24
UW-SO	0.78 (0.00)	6.82 (0.17)	60	-10.10
GLS	0.79 (0.01)	6.73 (0.03)	44	-11.26
IMTL	0.78 (0.00)	6.81 (0.11)	40	-10.46
GradNorm	0.78 (0.00)	6.77 (0.06)	47	-10.70
NashMTL	0.77 (0.01)	6.84 (0.18)	55	-9.62
GradDrop	0.77 (0.01)	6.99 (0.18)	46	-8.25
PCGrad	0.78 (0.00)	6.86 (0.10)	49	-9.83
CAGrad	0.78 (0.01)	6.88 (0.20)	125	-9.68

Table 35: Results on the Helpdesk dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.66 (0.01)	5.58 (0.07)	/	/
EW	0.78 (0.01)	5.52 (0.02)	54	-9.02
Scalarization	0.79 (0.00)	5.39 (0.12)	39	-10.94
DWA	0.77 (0.00)	5.53 (0.08)	41	-8.81
RLW	0.77 (0.01)	5.62 (0.05)	45	-7.89
UW	0.78 (0.00)	5.36 (0.12)	52	-11.02
UW-O	0.79 (0.00)	5.41 (0.12)	45	-10.80
UW-SO	0.78 (0.00)	5.45 (0.11)	51	-10.19
GLS	0.78 (0.00)	5.39 (0.10)	46	-10.83
IMTL	0.78 (0.00)	5.49 (0.07)	30	-9.99
GradNorm	0.78 (0.00)	5.38 (0.17)	56	-10.93
NashMTL	0.77 (0.00)	5.62 (0.10)	121	-7.79
GradDrop	0.78 (0.01)	5.53 (0.04)	51	-8.98
PCGrad	0.78 (0.01)	5.45 (0.08)	49	-9.78
CAGrad	0.79 (0.01)	5.39 (0.08)	48	-10.95

Table 36: Results on the Helpdesk dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.68 (0.01)	5.07 (0.07)	6.31 (0.15)	/	/
EW	0.80 (0.01)	5.10 (0.08)	6.29 (0.04)	25	-5.89
DWA	0.80 (0.01)	5.13 (0.13)	6.28 (0.01)	28	-5.88
RLW	0.80 (0.01)	5.13 (0.16)	6.37 (0.07)	27	-5.27
UW	0.80 (0.01)	5.14 (0.11)	6.33 (0.06)	26	-5.69
UW-O	0.81 (0.00)	5.08 (0.08)	6.31 (0.05)	22	-6.48
UW-SO	0.80 (0.00)	5.07 (0.15)	6.28 (0.01)	35	-6.44
GLS	0.81 (0.00)	5.10 (0.11)	6.39 (0.07)	27	-5.90
IMTL	0.80 (0.00)	5.16 (0.11)	6.55 (0.02)	26	-4.34
GradNorm	0.81 (0.00)	5.10 (0.09)	6.32 (0.03)	29	-6.21
NashMTL	0.80 (0.01)	5.15 (0.17)	6.34 (0.19)	19	-5.26
GradDrop	0.79 (0.01)	5.12 (0.09)	6.34 (0.03)	27	-5.27
PCGrad	0.80 (0.00)	5.11 (0.16)	6.29 (0.07)	31	-6.28
CAGrad	0.81 (0.00)	5.14 (0.18)	6.30 (0.07)	30	-6.16

Table 37: Results on the Helpdesk dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.68 (0.01)	5.07 (0.07)	/	/
EW	0.80 (0.01)	5.06 (0.08)	21	-9.55
Scalarization	0.81 (0.01)	5.13 (0.15)	30	-9.05
DWA	0.80 (0.00)	5.12 (0.18)	19	-8.77
RLW	0.80 (0.01)	5.10 (0.08)	27	-8.87
UW	0.80 (0.02)	5.14 (0.10)	22	-8.31
UW-O	0.80 (0.01)	5.20 (0.11)	18	-8.06
UW-SO	0.81 (0.01)	5.08 (0.09)	26	-9.58
GLS	0.81 (0.00)	5.15 (0.16)	22	-8.86
IMTL	0.81 (0.00)	5.21 (0.21)	28	-8.45
GradNorm	0.81 (0.00)	5.07 (0.12)	24	-9.77
NashMTL	0.80 (0.01)	5.09 (0.14)	23	-8.88
GradDrop	0.80 (0.00)	5.11 (0.17)	26	-8.99
PCGrad	0.81 (0.00)	5.13 (0.09)	26	-9.06
CAGrad	0.81 (0.00)	5.09 (0.11)	33	-9.79

Table 38: Results on the Helpdesk dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	5.07 (0.07)	6.31 (0.15)	/	/
EW	5.11 (0.08)	6.33 (0.07)	20	0.54
DWA	5.12 (0.03)	6.33 (0.07)	20	0.58
RLW	5.10 (0.10)	6.32 (0.08)	33	0.27
UW	5.11 (0.19)	6.32 (0.07)	24	0.43
UW-O	5.10 (0.03)	6.38 (0.14)	20	0.84
UW-SO	5.06 (0.11)	6.27 (0.07)	28	-0.43
GLS	5.11 (0.07)	6.34 (0.08)	21	0.60
IMTL	5.35 (0.10)	6.50 (0.08)	12	4.18
GradNorm	5.07 (0.10)	6.38 (0.15)	25	0.47
NashMTL	5.12 (0.10)	6.30 (0.13)	27	0.31
GradDrop	5.03 (0.14)	6.30 (0.08)	30	-0.56
PCGrad	5.14 (0.18)	6.31 (0.08)	20	0.70
CAGrad	5.09 (0.06)	6.33 (0.04)	28	0.22

Table 39: Results on the Sepsis dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.48 (0.01)	1.55 (0.02)	19.28 (1.42)	/	/
EW	0.58 (0.01)	1.18 (0.14)	18.58 (1.35)	100	-16.09
DWA	0.58 (0.02)	1.16 (0.09)	18.97 (1.39)	95	-15.99
RLW	0.58 (0.00)	1.23 (0.08)	19.25 (1.76)	94	-14.20
UW	0.60 (0.01)	1.62 (0.04)	23.21 (0.78)	72	-0.55
UW-O	0.60 (0.00)	1.49 (0.06)	22.05 (1.71)	94	-5.28
UW-SO	0.58 (0.01)	1.14 (0.11)	18.78 (1.25)	93	-16.86
GLS	0.61 (0.01)	1.56 (0.02)	22.55 (2.07)	110	-3.26
IMTL	0.60 (0.00)	1.49 (0.12)	20.32 (1.01)	82	-7.85
GradNorm	0.58 (0.01)	1.26 (0.07)	18.71 (0.81)	99	-14.51
NashMTL	0.58 (0.01)	1.10 (0.10)	18.66 (1.80)	81	-17.92
GradDrop	0.57 (0.01)	1.17 (0.05)	18.88 (1.48)	99	-15.22
PCGrad	0.59 (0.01)	1.25 (0.16)	18.93 (1.57)	84	-14.89
CAGrad	0.59 (0.02)	1.26 (0.11)	19.10 (1.22)	80	-14.63

Table 40: Results on the Sepsis dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.48 (0.01)	1.55 (0.02)	/	/
EW	0.60 (0.02)	1.61 (0.03)	12	-10.73
DWA	0.60 (0.01)	1.57 (0.02)	10	-12.85
RLW	0.59 (0.03)	1.58 (0.07)	11	-10.96
UW	0.60 (0.01)	1.59 (0.01)	10	-11.61
UW-O	0.61 (0.02)	1.63 (0.05)	13	-11.66
UW-SO	0.61 (0.01)	1.58 (0.08)	10	-12.66
GLS	0.61 (0.01)	1.61 (0.05)	22	-11.98
IMTL	0.62 (0.01)	1.56 (0.04)	19	-14.42
GradNorm	0.60 (0.02)	1.63 (0.03)	10	-10.80
NashMTL	0.61 (0.01)	1.59 (0.06)	16	-12.90
GradDrop	0.60 (0.01)	1.61 (0.08)	17	-10.86
PCGrad	0.60 (0.01)	1.55 (0.01)	9	-13.39
CAGrad	0.60 (0.02)	1.57 (0.05)	11	-12.60

Table 41: Results on the Sepsis dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.48 (0.01)	19.28 (1.42)	/	/
EW	0.57 (0.01)	18.71 (1.38)	85	-11.29
DWA	0.58 (0.01)	18.95 (1.16)	78	-11.22
RLW	0.58 (0.02)	19.19 (1.03)	76	-10.85
UW	0.59 (0.01)	20.69 (1.03)	65	-7.72
UW-O	0.60 (0.01)	20.26 (0.36)	102	-10.43
UW-SO	0.57 (0.01)	18.78 (1.23)	82	-11.02
GLS	0.60 (0.01)	20.43 (1.60)	95	-9.93
IMTL	0.60 (0.01)	19.89 (0.62)	79	-11.14
GradNorm	0.59 (0.01)	19.03 (1.53)	83	-12.13
NashMTL	0.57 (0.01)	18.63 (1.57)	83	-11.78
GradDrop	0.57 (0.01)	19.17 (1.22)	72	-10.35
PCGrad	0.59 (0.01)	18.88 (1.46)	81	-12.57
CAGrad	0.59 (0.00)	18.71 (0.93)	74	-13.19

Table 42: Results on the Sepsis dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.55 (0.02)	19.28 (1.42)	/	/
EW	1.24 (0.14)	20.06 (1.58)	85	-8.19
DWA	1.18 (0.07)	19.30 (1.32)	82	-11.90
RLW	1.23 (0.24)	19.21 (1.41)	129	-10.70
UW	1.78 (0.15)	24.71 (1.36)	112	21.26
UW-O	1.54 (0.10)	23.86 (0.93)	73	11.55
UW-SO	1.20 (0.07)	19.05 (1.09)	111	-12.05
GLS	1.57 (0.09)	25.00 (2.41)	79	15.18
IMTL	1.36 (0.12)	19.83 (0.99)	108	-4.88
GradNorm	1.22 (0.12)	19.53 (1.67)	95	-10.06
NashMTL	1.05 (0.16)	18.57 (1.28)	118	-17.97
GradDrop	1.19 (0.08)	19.41 (1.32)	78	-11.50
PCGrad	1.10 (0.20)	18.84 (1.17)	103	-15.70
CAGrad	1.10 (0.11)	18.70 (1.02)	116	-15.98

Table 43: Results on the Sepsis dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.43 (0.02)	1.90 (0.10)	23.94 (1.50)	/	/
EW	0.52 (0.01)	1.91 (0.11)	24.44 (0.64)	93	-6.41
DWA	0.50 (0.02)	1.98 (0.18)	26.34 (1.70)	52	-0.86
RLW	0.51 (0.01)	1.95 (0.11)	25.62 (1.55)	62	-2.92
UW	0.55 (0.02)	1.94 (0.15)	26.61 (0.40)	48	-4.85
UW-O	0.58 (0.01)	1.88 (0.16)	26.90 (1.01)	49	-7.50
UW-SO	0.52 (0.02)	1.97 (0.11)	24.61 (1.10)	91	-5.02
GLS	0.59 (0.01)	1.87 (0.13)	27.43 (1.08)	24	-7.68
IMTL	0.58 (0.01)	1.85 (0.14)	26.57 (1.10)	47	-8.69
GradNorm	0.55 (0.01)	1.95 (0.10)	25.98 (0.99)	113	-5.25
NashMTL	0.52 (0.01)	2.01 (0.11)	24.96 (0.81)	149	-3.80
GradDrop	0.49 (0.01)	1.97 (0.11)	23.88 (1.07)	85	-3.60
PCGrad	0.54 (0.02)	1.87 (0.04)	24.67 (1.00)	95	-7.85
CAGrad	0.56 (0.01)	1.91 (0.11)	24.97 (0.96)	88	-8.77

Table 44: Results on the Sepsis dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.90 (0.10)	23.94 (1.50)	/	/
EW	1.98 (0.15)	25.01 (0.66)	58	4.48
DWA	2.10 (0.26)	23.93 (1.08)	93	5.35
RLW	2.01 (0.09)	24.60 (0.95)	83	4.32
UW	1.99 (0.11)	26.10 (1.24)	100	6.95
UW-O	1.94 (0.13)	26.55 (0.64)	49	6.46
UW-SO	1.99 (0.13)	24.07 (1.34)	76	2.68
GLS	1.98 (0.13)	26.99 (0.57)	74	8.41
IMTL	1.96 (0.03)	26.14 (1.17)	58	6.32
GradNorm	2.02 (0.13)	24.62 (1.39)	68	4.50
NashMTL	2.02 (0.13)	25.77 (0.89)	108	7.01
GradDrop	2.02 (0.16)	24.83 (0.80)	80	4.95
PCGrad	1.92 (0.09)	24.40 (1.25)	94	1.44
CAGrad	1.93 (0.16)	24.21 (1.14)	81	1.33

Table 45: Results on the Sepsis dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.43 (0.02)	23.94 (1.50)	/	/
EW	0.50 (0.01)	24.00 (0.99)	79	-8.37
DWA	0.51 (0.00)	24.21 (1.22)	62	-8.54
RLW	0.52 (0.01)	25.25 (1.46)	49	-7.39
UW	0.57 (0.01)	26.58 (0.80)	38	-10.95
UW-O	0.58 (0.01)	27.50 (1.00)	27	-10.39
UW-SO	0.50 (0.01)	23.75 (1.61)	82	-9.07
GLS	0.57 (0.02)	27.25 (1.20)	78	-9.34
IMTL	0.58 (0.01)	26.17 (1.09)	36	-12.41
GradNorm	0.56 (0.01)	25.04 (1.38)	60	-13.22
NashMTL	0.53 (0.01)	25.35 (0.96)	134	-8.88
GradDrop	0.49 (0.03)	26.00 (1.84)	54	-2.54
PCGrad	0.52 (0.00)	24.34 (0.97)	66	-9.71
CAGrad	0.55 (0.00)	24.58 (1.12)	55	-12.69

Table 46: Results on the Sepsis dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.43 (0.02)	1.90 (0.10)	/	/
EW	0.57 (0.01)	1.86 (0.07)	64	-17.00
DWA	0.57 (0.00)	1.85 (0.08)	69	-17.08
RLW	0.57 (0.01)	1.89 (0.08)	63	-16.22
UW	0.58 (0.01)	1.87 (0.12)	16	-17.79
UW-O	0.58 (0.01)	1.87 (0.10)	17	-18.19
UW-SO	0.58 (0.00)	1.86 (0.08)	17	-17.83
GLS	0.58 (0.00)	1.81 (0.05)	101	-19.24
IMTL	0.57 (0.02)	1.80 (0.04)	11	-19.04
GradNorm	0.57 (0.01)	1.87 (0.10)	66	-16.98
NashMTL	0.58 (0.02)	1.80 (0.04)	10	-19.95
GradDrop	0.57 (0.01)	1.87 (0.09)	69	-17.11
PCGrad	0.57 (0.00)	1.85 (0.13)	10	-17.39
CAGrad	0.58 (0.01)	1.87 (0.09)	15	-17.83

Table 47: Results on the Sepsis dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.43 (0.09)	1.69 (0.07)	24.97 (1.44)	/	/
EW	0.53 (0.12)	1.76 (0.07)	25.47 (2.31)	42	-5.33
DWA	0.61 (0.01)	1.99 (0.37)	24.75 (2.01)	57	-7.92
RLW	0.60 (0.04)	1.69 (0.27)	25.29 (1.95)	66	-12.35
UW	0.61 (0.01)	1.69 (0.09)	25.18 (1.39)	55	-13.55
UW-O	0.64 (0.01)	1.70 (0.09)	27.12 (0.99)	55	-13.24
UW-SO	0.61 (0.02)	1.73 (0.16)	24.64 (2.08)	58	-12.89
GLS	0.64 (0.02)	1.66 (0.07)	27.25 (0.67)	21	-13.03
IMTL	0.64 (0.01)	1.58 (0.03)	27.15 (0.94)	22	-15.10
GradNorm	0.62 (0.01)	1.77 (0.04)	25.32 (1.51)	53	-12.50
NashMTL	0.61 (0.01)	1.83 (0.09)	25.29 (1.89)	42	-10.67
GradDrop	0.49 (0.05)	1.69 (0.08)	25.40 (1.52)	45	-4.14
PCGrad	0.62 (0.01)	1.88 (0.17)	24.89 (1.12)	62	-10.57
CAGrad	0.63 (0.01)	1.65 (0.04)	25.65 (0.70)	73	-14.65

Table 48: Results on the Sepsis dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.69 (0.07)	24.97 (1.44)	/	/
EW	1.73 (0.16)	29.75 (6.26)	41	10.93
DWA	1.70 (0.08)	29.58 (3.13)	35	9.66
RLW	1.81 (0.03)	25.11 (1.67)	60	3.82
UW	1.75 (0.14)	27.29 (1.47)	64	6.45
UW-O	1.60 (0.06)	26.51 (1.23)	43	0.39
UW-SO	1.82 (0.07)	25.21 (1.61)	58	4.50
GLS	1.68 (0.04)	26.57 (1.45)	47	3.10
IMTL	1.72 (0.03)	26.31 (1.78)	37	3.57
GradNorm	1.71 (0.10)	25.19 (1.53)	61	1.10
NashMTL	1.62 (0.03)	26.18 (1.23)	30	0.51
GradDrop	1.87 (0.27)	25.78 (1.62)	53	7.02
PCGrad	1.77 (0.14)	25.27 (1.22)	46	2.94
CAGrad	1.95 (0.10)	24.81 (3.21)	71	7.46

Table 49: Results on the Sepsis dataset for Transformer with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.43 (0.09)	24.97 (1.44)	/	/
EW	0.60 (0.02)	24.49 (2.43)	66	-20.77
DWA	0.61 (0.01)	25.07 (1.75)	66	-19.79
RLW	0.60 (0.01)	25.34 (1.00)	56	-18.60
UW	0.61 (0.02)	25.49 (2.24)	33	-19.45
UW-O	0.64 (0.00)	26.64 (1.23)	40	-20.31
UW-SO	0.61 (0.01)	24.75 (2.59)	63	-20.28
GLS	0.64 (0.02)	26.44 (1.28)	47	-20.92
IMTL	0.64 (0.01)	26.37 (1.13)	35	-21.29
GradNorm	0.63 (0.01)	25.41 (1.49)	54	-21.93
NashMTL	0.60 (0.02)	25.43 (1.80)	45	-18.04
GradDrop	0.59 (0.02)	25.46 (2.03)	44	-16.56
PCGrad	0.62 (0.00)	24.85 (1.50)	56	-21.82
CAGrad	0.63 (0.02)	25.16 (1.18)	61	-21.78

Table 50: Results on the Sepsis dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.43 (0.09)	1.69 (0.07)	/	/
EW	0.63 (0.02)	1.58 (0.02)	26	-26.26
DWA	0.64 (0.01)	1.63 (0.07)	24	-25.89
RLW	0.64 (0.02)	1.58 (0.02)	17	-27.11
UW	0.64 (0.02)	1.54 (0.03)	26	-28.58
UW-O	0.65 (0.01)	1.63 (0.06)	26	-26.30
UW-SO	0.64 (0.02)	1.64 (0.13)	28	-25.66
GLS	0.64 (0.01)	1.59 (0.03)	25	-26.56
IMTL	0.64 (0.01)	1.58 (0.02)	18	-26.86
GradNorm	0.64 (0.02)	1.58 (0.04)	23	-27.14
NashMTL	0.64 (0.02)	1.63 (0.06)	20	-25.24
GradDrop	0.63 (0.01)	1.63 (0.07)	17	-24.61
PCGrad	0.63 (0.01)	1.55 (0.07)	23	-27.23
CAGrad	0.63 (0.01)	1.54 (0.02)	29	-27.23

Table 51: Results on the BPIC20DD dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.71 (0.01)	1.64 (0.03)	2.80 (0.02)	/	/
EW	0.86 (0.01)	1.64 (0.03)	2.79 (0.03)	52	-7.20
DWA	0.86 (0.00)	1.64 (0.02)	2.78 (0.03)	51	-7.25
RLW	0.86 (0.01)	1.64 (0.03)	2.79 (0.01)	57	-7.09
UW	0.86 (0.01)	1.65 (0.02)	2.81 (0.03)	31	-6.71
UW-O	0.86 (0.01)	1.65 (0.01)	2.81 (0.03)	40	-6.61
UW-SO	0.86 (0.01)	1.64 (0.03)	2.78 (0.03)	50	-7.19
GLS	0.86 (0.01)	1.65 (0.02)	2.82 (0.02)	37	-6.58
IMTL	0.86 (0.01)	1.65 (0.03)	2.80 (0.02)	52	-6.81
GradNorm	0.86 (0.01)	1.65 (0.03)	2.79 (0.01)	48	-6.88
NashMTL	0.86 (0.01)	1.64 (0.03)	2.79 (0.02)	40	-7.07
GradDrop	0.86 (0.01)	1.64 (0.03)	2.79 (0.01)	43	-7.02
PCGrad	0.86 (0.01)	1.65 (0.03)	2.80 (0.02)	41	-6.93
CAGrad	0.86 (0.01)	1.65 (0.03)	2.80 (0.03)	38	-6.85

Table 52: Results on the BPIC20DD dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.71 (0.01)	1.64 (0.03)	/	/
EW	0.86 (0.01)	1.64 (0.04)	30	-10.60
DWA	0.86 (0.01)	1.64 (0.03)	30	-10.47
RLW	0.86 (0.01)	1.65 (0.02)	31	-10.37
UW	0.86 (0.01)	1.65 (0.02)	32	-10.21
UW-O	0.86 (0.01)	1.65 (0.02)	33	-10.24
UW-SO	0.86 (0.01)	1.64 (0.03)	27	-10.48
GLS	0.86 (0.01)	1.65 (0.02)	28	-10.14
IMTL	0.86 (0.01)	1.66 (0.02)	126	-9.88
GradNorm	0.86 (0.01)	1.65 (0.03)	34	-10.16
NashMTL	0.86 (0.01)	1.65 (0.03)	23	-10.32
GradDrop	0.86 (0.01)	1.64 (0.02)	33	-10.52
PCGrad	0.86 (0.01)	1.65 (0.02)	32	-10.36
CAGrad	0.86 (0.01)	1.65 (0.02)	27	-10.31

Table 53: Results on the BPIC20DD dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.71 (0.01)	2.80 (0.02)	/	/
EW	0.86 (0.01)	2.81 (0.03)	41	-10.38
DWA	0.86 (0.01)	2.80 (0.02)	42	-10.52
RLW	0.86 (0.01)	2.82 (0.01)	35	-10.22
UW	0.86 (0.01)	2.82 (0.03)	25	-10.13
UW-O	0.86 (0.01)	2.83 (0.02)	22	-9.90
UW-SO	0.86 (0.01)	2.80 (0.02)	42	-10.48
GLS	0.86 (0.01)	2.83 (0.03)	21	-9.99
IMTL	0.86 (0.01)	2.83 (0.03)	25	-10.02
GradNorm	0.86 (0.01)	2.81 (0.01)	33	-10.31
NashMTL	0.86 (0.01)	2.80 (0.02)	34	-10.46
GradDrop	0.86 (0.01)	2.83 (0.02)	21	-10.09
PCGrad	0.86 (0.01)	2.81 (0.02)	32	-10.40
CAGrad	0.86 (0.01)	2.82 (0.03)	21	-10.13

Table 54: Results on the BPIC20DD dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.64 (0.03)	2.80 (0.02)	/	/
EW	1.63 (0.03)	2.77 (0.04)	62	-0.82
DWA	1.63 (0.04)	2.77 (0.04)	53	-0.81
RLW	1.63 (0.03)	2.76 (0.04)	68	-0.97
UW	1.63 (0.03)	2.77 (0.02)	46	-0.73
UW-O	1.63 (0.03)	2.78 (0.03)	53	-0.66
UW-SO	1.63 (0.03)	2.76 (0.04)	65	-1.02
GLS	1.63 (0.03)	2.77 (0.03)	59	-0.86
IMTL	1.64 (0.03)	2.78 (0.03)	37	-0.34
GradNorm	1.63 (0.03)	2.77 (0.03)	59	-0.87
NashMTL	1.63 (0.03)	2.77 (0.03)	63	-0.77
GradDrop	1.63 (0.03)	2.78 (0.03)	52	-0.61
PCGrad	1.64 (0.03)	2.77 (0.04)	57	-0.72
CAGrad	1.63 (0.04)	2.77 (0.03)	55	-0.83

Table 55: Results on the BPIC20DD dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.67 (0.01)	1.68 (0.01)	2.86 (0.04)	/	/
EW	0.86 (0.00)	1.67 (0.03)	2.82 (0.02)	41	-9.79
DWA	0.86 (0.01)	1.67 (0.03)	2.83 (0.02)	33	-9.62
RLW	0.86 (0.01)	1.68 (0.02)	2.84 (0.03)	36	-9.45
UW	0.86 (0.01)	1.68 (0.03)	2.85 (0.03)	38	-9.21
UW-O	0.86 (0.00)	1.69 (0.02)	2.84 (0.02)	33	-9.22
UW-SO	0.86 (0.01)	1.67 (0.03)	2.83 (0.02)	46	-9.77
GLS	0.86 (0.01)	1.69 (0.02)	2.84 (0.02)	35	-9.28
IMTL	0.86 (0.01)	1.68 (0.03)	2.85 (0.02)	34	-9.19
GradNorm	0.86 (0.00)	1.67 (0.02)	2.83 (0.02)	38	-9.65
NashMTL	0.86 (0.00)	1.68 (0.04)	2.83 (0.02)	30	-9.57
GradDrop	0.86 (0.01)	1.68 (0.01)	2.83 (0.03)	39	-9.53
PCGrad	0.86 (0.01)	1.67 (0.03)	2.83 (0.02)	43	-9.58
CAGrad	0.86 (0.01)	1.67 (0.03)	2.83 (0.03)	47	-9.63

Table 56: Results on the BPIC20DD dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.68 (0.01)	2.86 (0.04)	/	/
EW	1.67 (0.04)	2.83 (0.02)	36	-0.85
DWA	1.66 (0.03)	2.83 (0.03)	43	-1.07
RLW	1.68 (0.02)	2.84 (0.03)	32	-0.52
UW	1.67 (0.02)	2.84 (0.02)	29	-0.68
UW-O	1.67 (0.03)	2.83 (0.02)	29	-0.92
UW-SO	1.66 (0.04)	2.82 (0.02)	54	-1.31
GLS	1.66 (0.04)	2.83 (0.02)	34	-1.05
IMTL	1.66 (0.04)	2.83 (0.03)	42	-1.27
GradNorm	1.67 (0.03)	2.83 (0.02)	44	-0.95
NashMTL	1.66 (0.03)	2.83 (0.03)	35	-0.94
GradDrop	1.67 (0.02)	2.83 (0.02)	48	-0.74
PCGrad	1.66 (0.03)	2.83 (0.02)	39	-1.03
CAGrad	1.66 (0.02)	2.83 (0.02)	48	-1.02

Table 57: Results on the BPIC20DD dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.67 (0.01)	2.86 (0.04)	/	/
EW	0.86 (0.01)	2.86 (0.02)	6	-13.70
DWA	0.86 (0.01)	2.85 (0.02)	7	-13.73
RLW	0.86 (0.00)	2.86 (0.03)	11	-13.73
UW	0.86 (0.00)	2.85 (0.02)	26	-13.92
UW-O	0.86 (0.01)	2.85 (0.02)	12	-13.95
UW-SO	0.86 (0.01)	2.85 (0.03)	27	-13.95
GLS	0.86 (0.00)	2.85 (0.03)	21	-13.93
IMTL	0.86 (0.01)	2.85 (0.03)	19	-13.91
GradNorm	0.86 (0.01)	2.86 (0.03)	12	-13.85
NashMTL	0.86 (0.01)	2.83 (0.02)	27	-14.22
GradDrop	0.86 (0.01)	2.85 (0.03)	14	-13.94
PCGrad	0.86 (0.01)	2.85 (0.03)	6	-13.77
CAGrad	0.86 (0.01)	2.85 (0.03)	26	-13.95

Table 58: Results on the BPIC20DD dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.67 (0.01)	1.68 (0.01)	/	/
EW	0.86 (0.01)	1.66 (0.02)	45	-14.17
DWA	0.86 (0.01)	1.67 (0.02)	31	-14.09
RLW	0.86 (0.01)	1.67 (0.04)	42	-14.09
UW	0.86 (0.01)	1.68 (0.02)	42	-13.59
UW-O	0.86 (0.01)	1.69 (0.03)	29	-13.42
UW-SO	0.86 (0.01)	1.66 (0.03)	47	-14.18
GLS	0.86 (0.01)	1.68 (0.03)	34	-13.57
IMTL	0.86 (0.00)	1.68 (0.03)	29	-13.55
GradNorm	0.86 (0.00)	1.69 (0.04)	33	-13.48
NashMTL	0.86 (0.01)	1.66 (0.02)	44	-14.35
GradDrop	0.86 (0.01)	1.68 (0.02)	28	-13.54
PCGrad	0.86 (0.01)	1.68 (0.04)	26	-13.63
CAGrad	0.86 (0.01)	1.67 (0.03)	31	-14.04

Table 59: Results on the BPIC20DD dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.83 (0.01)	1.67 (0.02)	2.85 (0.04)	/	/
EW	0.86 (0.01)	1.68 (0.02)	2.84 (0.02)	48	-0.97
DWA	0.86 (0.01)	1.69 (0.02)	2.84 (0.03)	31	-0.82
RLW	0.86 (0.01)	1.68 (0.03)	2.84 (0.03)	31	-0.93
UW	0.86 (0.01)	1.68 (0.02)	2.84 (0.02)	36	-0.90
UW-O	0.86 (0.01)	1.69 (0.02)	2.85 (0.03)	32	-0.70
UW-SO	0.86 (0.01)	1.68 (0.03)	2.84 (0.03)	41	-0.93
GLS	0.86 (0.01)	1.68 (0.02)	2.85 (0.04)	38	-0.79
IMTL	0.86 (0.01)	1.68 (0.02)	2.85 (0.03)	31	-0.93
GradNorm	0.86 (0.01)	1.68 (0.03)	2.85 (0.02)	36	-0.92
NashMTL	0.86 (0.01)	1.68 (0.03)	2.84 (0.03)	42	-1.02
GradDrop	0.86 (0.01)	1.68 (0.02)	2.84 (0.03)	27	-0.95
PCGrad	0.86 (0.01)	1.68 (0.02)	2.84 (0.02)	27	-0.91
CAGrad	0.85 (0.00)	1.69 (0.01)	2.88 (0.03)	18	-0.20

Table 60: Results on the BPIC20DD dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.83 (0.01)	1.67 (0.02)	/	/
EW	0.86 (0.01)	1.68 (0.02)	30	-1.39
DWA	0.86 (0.01)	1.68 (0.02)	38	-1.49
RLW	0.86 (0.01)	1.68 (0.03)	28	-1.29
UW	0.86 (0.01)	1.67 (0.02)	39	-1.61
UW-O	0.86 (0.01)	1.68 (0.03)	33	-1.44
UW-SO	0.86 (0.01)	1.68 (0.02)	39	-1.48
GLS	0.86 (0.01)	1.68 (0.02)	40	-1.35
IMTL	0.86 (0.01)	1.68 (0.02)	33	-1.41
GradNorm	0.86 (0.01)	1.68 (0.02)	32	-1.54
NashMTL	0.86 (0.01)	1.68 (0.02)	36	-1.39
GradDrop	0.86 (0.01)	1.68 (0.03)	25	-1.42
PCGrad	0.86 (0.01)	1.68 (0.02)	29	-1.45
CAGrad	0.86 (0.01)	1.68 (0.02)	41	-1.38

Table 61: Results on the BPIC20DD dataset for Transformer with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.83 (0.01)	2.85 (0.04)	/	/
EW	0.86 (0.01)	2.84 (0.03)	35	-1.72
DWA	0.86 (0.01)	2.84 (0.03)	31	-1.64
RLW	0.86 (0.01)	2.84 (0.03)	28	-1.67
UW	0.86 (0.01)	2.85 (0.02)	24	-1.60
UW-O	0.86 (0.01)	2.85 (0.03)	27	-1.58
UW-SO	0.86 (0.01)	2.84 (0.03)	32	-1.74
GLS	0.86 (0.01)	2.84 (0.02)	32	-1.69
IMTL	0.86 (0.01)	2.87 (0.01)	13	-1.21
GradNorm	0.86 (0.01)	2.85 (0.03)	36	-1.54
NashMTL	0.86 (0.01)	2.84 (0.02)	33	-1.76
GradDrop	0.86 (0.01)	2.85 (0.03)	21	-1.53
PCGrad	0.86 (0.01)	2.84 (0.02)	23	-1.81
CAGrad	0.86 (0.01)	2.85 (0.03)	29	-1.68

Table 62: Results on the BPIC20DD dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.67 (0.02)	2.85 (0.04)	/	/
EW	1.69 (0.03)	2.84 (0.03)	31	0.34
DWA	1.68 (0.02)	2.84 (0.03)	29	0.24
RLW	1.68 (0.02)	2.84 (0.03)	31	0.16
UW	1.68 (0.02)	2.83 (0.02)	39	0.16
UW-O	1.68 (0.02)	2.84 (0.02)	30	0.08
UW-SO	1.68 (0.03)	2.83 (0.02)	50	-0.03
GLS	1.68 (0.03)	2.84 (0.03)	33	0.21
IMTL	1.68 (0.02)	2.84 (0.02)	31	0.16
GradNorm	1.68 (0.03)	2.83 (0.02)	48	0.01
NashMTL	1.68 (0.02)	2.83 (0.03)	38	0.18
GradDrop	1.69 (0.02)	2.84 (0.03)	29	0.45
PCGrad	1.68 (0.03)	2.83 (0.02)	43	-0.02
CAGrad	1.68 (0.03)	2.82 (0.02)	53	-0.13

Table 63: Results on the BPIC20ID dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.62 (0.00)	5.72 (0.04)	15.48 (0.30)	/	/
EW	0.85 (0.00)	5.60 (0.05)	15.54 (0.20)	70	-12.90
Scalarization	0.85 (0.00)	5.64 (0.05)	15.75 (0.10)	187	-12.40
DWA	0.85 (0.00)	5.58 (0.02)	15.47 (0.01)	74	-13.10
RLW	0.85 (0.00)	5.59 (0.06)	15.72 (0.10)	65	-12.53
UW	0.85 (0.01)	5.66 (0.04)	15.77 (0.11)	57	-12.25
UW-O	0.86 (0.00)	5.73 (0.05)	16.06 (0.19)	41	-11.51
UW-SO	0.85 (0.00)	5.54 (0.05)	15.40 (0.10)	82	-13.65
GLS	0.86 (0.00)	5.65 (0.05)	16.05 (0.14)	49	-11.89
IMTL	0.85 (0.00)	5.66 (0.04)	15.75 (0.11)	70	-12.34
GradNorm	0.85 (0.00)	5.58 (0.10)	15.45 (0.26)	83	-13.32
NashMTL	0.85 (0.00)	5.53 (0.02)	15.39 (0.17)	96	-13.59
GradDrop	0.85 (0.00)	5.59 (0.04)	15.72 (0.08)	58	-12.41
PCGrad	0.86 (0.00)	5.54 (0.07)	15.56 (0.18)	70	-13.50
CAGrad	0.86 (0.00)	5.58 (0.04)	15.53 (0.08)	69	-13.30

Table 64: Results on the BPIC20ID dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.62 (0.00)	5.72 (0.04)	/	/
EW	0.85 (0.00)	5.74 (0.03)	29	-18.65
Scalarization	0.86 (0.00)	5.71 (0.02)	37	-19.39
DWA	0.86 (0.00)	5.72 (0.04)	28	-18.89
RLW	0.86 (0.00)	5.73 (0.04)	32	-19.04
UW	0.86 (0.00)	5.74 (0.04)	30	-19.18
UW-O	0.86 (0.00)	5.79 (0.07)	29	-18.90
UW-SO	0.86 (0.00)	5.69 (0.03)	33	-19.36
GLS	0.86 (0.00)	5.73 (0.04)	40	-19.15
IMTL	0.86 (0.00)	5.71 (0.03)	31	-19.43
GradNorm	0.86 (0.00)	5.73 (0.04)	26	-19.23
NashMTL	0.86 (0.00)	5.72 (0.07)	34	-19.06
GradDrop	0.86 (0.00)	5.71 (0.05)	36	-19.10
PCGrad	0.86 (0.00)	5.72 (0.03)	42	-19.36
CAGrad	0.86 (0.00)	5.69 (0.05)	36	-19.68

Table 65: Results on the BPIC20ID dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.62 (0.00)	15.48 (0.30)	/	/
EW	0.85 (0.00)	15.58 (0.15)	77	-17.85
Scalarization	0.85 (0.00)	15.57 (0.09)	73	-18.36
DWA	0.85 (0.00)	15.64 (0.18)	63	-18.03
RLW	0.85 (0.00)	15.79 (0.15)	66	-17.51
UW	0.86 (0.00)	15.90 (0.07)	56	-17.65
UW-O	0.86 (0.01)	16.13 (0.30)	50	-16.97
UW-SO	0.85 (0.00)	15.59 (0.10)	85	-18.07
GLS	0.86 (0.00)	15.94 (0.13)	46	-17.50
IMTL	0.86 (0.00)	16.07 (0.10)	39	-17.27
GradNorm	0.85 (0.00)	15.54 (0.16)	73	-18.37
NashMTL	0.85 (0.00)	15.63 (0.10)	75	-17.92
GradDrop	0.84 (0.01)	15.46 (0.19)	104	-17.96
PCGrad	0.86 (0.00)	15.63 (0.08)	62	-18.39
CAGrad	0.85 (0.00)	15.58 (0.14)	68	-18.33

Table 66: Results on the BPIC20ID dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	5.72 (0.04)	15.48 (0.30)	/	/
EW	5.61 (0.05)	15.39 (0.12)	82	-1.20
DWA	5.55 (0.08)	15.35 (0.36)	91	-1.89
RLW	5.63 (0.04)	15.81 (0.14)	58	0.30
UW	5.65 (0.06)	15.45 (0.11)	95	-0.70
UW-O	5.66 (0.08)	15.75 (0.05)	70	0.34
UW-SO	5.58 (0.07)	15.38 (0.21)	91	-1.58
GLS	5.61 (0.02)	15.65 (0.43)	83	-0.37
IMTL	5.62 (0.07)	15.81 (0.29)	48	0.24
GradNorm	5.59 (0.05)	15.34 (0.13)	104	-1.55
NashMTL	5.61 (0.06)	15.59 (0.13)	64	-0.59
GradDrop	5.58 (0.00)	15.48 (0.01)	77	-1.18
PCGrad	5.59 (0.05)	15.34 (0.33)	101	-1.57
CAGrad	5.60 (0.06)	15.42 (0.06)	90	-1.25

Table 67: Results on the BPIC20ID dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.46 (0.00)	6.29 (0.16)	16.92 (0.38)	/	/
EW	0.80 (0.00)	6.08 (0.03)	17.14 (0.09)	90	-24.80
DWA	0.79 (0.01)	6.13 (0.04)	17.09 (0.16)	91	-24.39
RLW	0.78 (0.03)	6.17 (0.10)	17.48 (0.84)	93	-22.63
UW	0.85 (0.00)	6.05 (0.03)	17.30 (0.10)	102	-28.09
UW-O	0.85 (0.01)	6.09 (0.07)	17.65 (0.22)	89	-27.36
UW-SO	0.81 (0.01)	6.07 (0.06)	16.89 (0.20)	111	-25.93
GLS	0.85 (0.00)	6.08 (0.02)	17.65 (0.26)	78	-27.32
IMTL	0.84 (0.01)	6.14 (0.03)	17.78 (0.48)	60	-26.22
GradNorm	0.82 (0.01)	6.09 (0.16)	16.97 (0.07)	112	-26.43
NashMTL	0.78 (0.01)	6.25 (0.04)	17.97 (0.18)	50	-21.01
GradDrop	0.79 (0.01)	6.11 (0.03)	16.88 (0.22)	107	-24.62
PCGrad	0.82 (0.01)	6.02 (0.02)	17.01 (0.31)	102	-26.93
CAGrad	0.82 (0.01)	6.07 (0.09)	17.25 (0.55)	90	-26.26

Table 68: Results on the BPIC20ID dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	6.29 (0.16)	16.92 (0.38)	/	/
EW	6.19 (0.05)	17.53 (0.20)	58	1.00
UW-O	6.17 (0.08)	17.64 (0.25)	70	1.16
UW-SO	6.04 (0.01)	16.96 (0.09)	124	-1.93
CAGrad	6.00 (0.04)	17.19 (0.22)	95	-1.57

Table 69: Results on the BPIC20ID dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.46 (0.00)	16.92 (0.38)	/	/
EW	0.81 (0.01)	16.85 (0.07)	118	-37.56
DWA	0.80 (0.00)	17.10 (0.20)	87	-35.69
RLW	0.80 (0.01)	17.50 (0.10)	53	-34.78
UW	0.84 (0.00)	17.30 (0.09)	78	-39.99
UW-O	0.85 (0.00)	17.81 (0.31)	67	-38.70
UW-SO	0.82 (0.01)	16.82 (0.19)	117	-38.63
GLS	0.85 (0.01)	17.40 (0.43)	94	-40.15
IMTL	0.85 (0.00)	17.31 (0.17)	86	-40.43
GradNorm	0.84 (0.00)	17.05 (0.17)	98	-40.48
NashMTL	0.80 (0.01)	17.52 (0.17)	77	-34.83
GradDrop	0.79 (0.01)	17.23 (0.30)	81	-34.26
PCGrad	0.82 (0.01)	17.00 (0.21)	97	-38.34
CAGrad	0.84 (0.00)	17.07 (0.23)	93	-40.25

Table 70: Results on the BPIC20ID dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.46 (0.00)	6.29 (0.16)	/	/
EW	0.81 (0.01)	6.09 (0.03)	56	-38.69
Scalarization	0.84 (0.01)	6.04 (0.02)	67	-42.61
DWA	0.81 (0.00)	6.06 (0.09)	66	-39.49
RLW	0.82 (0.00)	6.10 (0.13)	79	-40.02
UW	0.84 (0.00)	6.03 (0.13)	72	-43.16
UW-O	0.84 (0.00)	6.20 (0.12)	52	-41.70
UW-SO	0.83 (0.00)	6.03 (0.14)	72	-41.26
GLS	0.84 (0.00)	6.11 (0.08)	46	-42.08
IMTL	0.85 (0.00)	6.06 (0.09)	66	-43.06
GradNorm	0.84 (0.00)	6.03 (0.09)	68	-42.65
NashMTL	0.82 (0.01)	6.16 (0.05)	47	-39.12
GradDrop	0.81 (0.00)	6.05 (0.08)	73	-38.85
PCGrad	0.83 (0.00)	6.00 (0.12)	82	-41.80
CAGrad	0.83 (0.01)	6.03 (0.16)	78	-41.86

Table 71: Results on the BPIC20ID dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.67 (0.09)	5.75 (0.06)	16.07 (0.20)	/	/
EW	0.85 (0.01)	5.82 (0.16)	16.29 (0.41)	31	-8.19
DWA	0.85 (0.01)	5.94 (0.18)	16.22 (0.38)	32	-7.66
RLW	0.85 (0.00)	5.92 (0.25)	16.55 (0.21)	31	-7.03
UW	0.87 (0.00)	5.85 (0.10)	16.67 (0.06)	24	-7.79
UW-O	0.87 (0.00)	5.87 (0.11)	16.88 (0.04)	23	-7.30
UW-SO	0.86 (0.00)	5.71 (0.11)	16.03 (0.20)	44	-9.66
GLS	0.87 (0.00)	5.81 (0.14)	16.64 (0.26)	28	-8.09
IMTL	0.87 (0.00)	5.73 (0.07)	16.57 (0.09)	34	-8.87
GradNorm	0.86 (0.00)	5.78 (0.02)	16.20 (0.42)	46	-9.15
NashMTL	0.86 (0.00)	5.87 (0.05)	16.19 (0.16)	30	-8.35
GradDrop	0.84 (0.01)	5.79 (0.08)	16.34 (0.16)	30	-7.37
PCGrad	0.86 (0.00)	5.75 (0.12)	16.35 (0.26)	33	-8.92
CAGrad	0.87 (0.00)	5.76 (0.15)	16.17 (0.36)	42	-9.37

Table 72: Results on the BPIC20ID dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.67 (0.09)	5.75 (0.06)	/	/
EW	0.86 (0.00)	5.77 (0.07)	23	-14.18
Scalarization	0.86 (0.00)	5.72 (0.10)	28	-14.59
DWA	0.86 (0.01)	5.73 (0.10)	25	-14.42
RLW	0.87 (0.00)	5.71 (0.05)	30	-14.79
UW	0.86 (0.01)	5.82 (0.08)	20	-13.30
UW-O	0.87 (0.00)	5.81 (0.09)	30	-14.17
UW-SO	0.87 (0.00)	5.76 (0.05)	26	-14.35
GLS	0.87 (0.00)	5.82 (0.10)	25	-14.10
IMTL	0.87 (0.00)	5.73 (0.03)	37	-14.71
GradNorm	0.86 (0.00)	5.69 (0.07)	31	-14.77
NashMTL	0.86 (0.00)	5.77 (0.12)	26	-14.15
GradDrop	0.86 (0.00)	5.67 (0.05)	38	-15.04
PCGrad	0.87 (0.00)	5.74 (0.03)	30	-14.55
CAGrad	0.86 (0.00)	5.73 (0.05)	37	-14.58

Table 73: Results on the BPIC20ID dataset for Transformer with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.67 (0.09)	16.07 (0.20)	/	/
EW	0.86 (0.00)	16.25 (0.09)	27	-13.33
DWA	0.86 (0.00)	16.26 (0.22)	36	-13.35
RLW	0.86 (0.01)	16.35 (0.15)	27	-12.82
UW	0.86 (0.00)	16.50 (0.10)	18	-13.07
UW-O	0.87 (0.00)	16.82 (0.23)	28	-12.25
UW-SO	0.86 (0.00)	16.06 (0.20)	37	-14.16
GLS	0.87 (0.00)	16.54 (0.30)	38	-13.12
IMTL	0.86 (0.01)	16.55 (0.08)	35	-12.77
GradNorm	0.86 (0.00)	16.15 (0.23)	36	-13.97
NashMTL	0.86 (0.00)	16.28 (0.10)	29	-13.27
GradDrop	0.85 (0.00)	16.28 (0.21)	26	-12.46
PCGrad	0.86 (0.00)	16.24 (0.07)	40	-13.81
CAGrad	0.86 (0.01)	16.22 (0.13)	55	-13.91

Table 74: Results on the BPIC20ID dataset for Transformer with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	5.75 (0.06)	16.07 (0.20)	/	/
UW-O	5.84 (0.09)	16.36 (0.11)	29	1.66
UW-SO	5.75 (0.05)	15.96 (0.14)	35	-0.33
CAGrad	5.78 (0.03)	16.22 (0.02)	26	0.77

Table 75: Results on the BPIC12C dataset for the LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.65 (0.01)	0.58 (0.01)	4.66 (0.12)	/	/
EW	0.79 (0.00)	0.59 (0.02)	5.21 (0.02)	200	-2.04
DWA	0.78 (0.00)	0.59 (0.01)	5.21 (0.01)	199	-1.90
RLW	0.78 (0.00)	0.59 (0.02)	5.38 (0.03)	198	-0.85
UW	0.78 (0.00)	0.60 (0.01)	5.84 (0.03)	200	2.89
UW-O	0.79 (0.00)	0.59 (0.02)	5.90 (0.05)	199	2.72
UW-SO	0.79 (0.00)	0.59 (0.02)	5.21 (0.00)	199	-2.03
GLS	0.79 (0.00)	0.59 (0.01)	5.88 (0.07)	197	2.65
IMTL	0.79 (0.00)	0.59 (0.01)	5.88 (0.05)	199	2.83
NashMTL	0.79 (0.00)	0.60 (0.01)	5.22 (0.02)	197	-1.46

Table 76: Results on the BPIC12C dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.65 (0.01)	0.58 (0.01)	/	/
EW	0.79 (0.00)	0.58 (0.01)	20	-10.42
DWA	0.79 (0.00)	0.58 (0.02)	66	-10.30
RLW	0.79 (0.00)	0.58 (0.01)	30	-10.57
UW	0.79 (0.00)	0.58 (0.02)	22	-10.73
UW-O	0.79 (0.00)	0.58 (0.01)	20	-10.11
UW-SO	0.79 (0.00)	0.58 (0.01)	21	-10.52
GLS	0.79 (0.01)	0.58 (0.01)	23	-9.91
IMTL	0.79 (0.00)	0.58 (0.01)	24	-10.72
GradNorm	0.79 (0.00)	0.57 (0.01)	27	-10.90
NashMTL	0.79 (0.00)	0.58 (0.01)	23	-10.78
GradDrop	0.79 (0.00)	0.58 (0.02)	27	-10.81
PCGrad	0.79 (0.00)	0.58 (0.01)	22	-10.37
CAGrad	0.79 (0.00)	0.57 (0.01)	27	-10.91

Table 77: Results on the BPIC12C dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.65 (0.01)	4.66 (0.12)	/	/
EW	0.78 (0.00)	5.22 (0.01)	199	-3.93
DWA	0.78 (0.00)	5.21 (0.04)	199	-4.05
RLW	0.78 (0.00)	5.34 (0.02)	199	-2.81
UW	0.78 (0.00)	5.66 (0.05)	191	1.01
UW-O	0.79 (0.00)	5.83 (0.06)	196	2.37
UW-SO	0.78 (0.00)	5.23 (0.06)	199	-3.92
GLS	0.78 (0.00)	5.79 (0.09)	199	2.06
IMTL	0.79 (0.00)	5.82 (0.04)	198	2.31
NashMTL	0.78 (0.00)	5.30 (0.03)	182	-3.18

Table 78: Results on the BPIC12C dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	0.58 (0.01)	4.66 (0.12)	/	/
EW	0.60 (0.01)	4.63 (0.00)	144	2.01
DWA	0.60 (0.01)	4.73 (0.09)	112	3.04
RLW	0.60 (0.02)	4.76 (0.11)	136	2.74
UW	0.61 (0.00)	5.38 (0.08)	91	10.28
UW-O	0.60 (0.01)	5.22 (0.06)	138	8.23
UW-SO	0.60 (0.02)	4.67 (0.08)	135	2.15
GLS	0.61 (0.02)	5.34 (0.08)	96	9.81
IMTL	0.61 (0.01)	4.95 (0.16)	117	5.85
GradNorm	0.62 (0.02)	4.83 (0.07)	127	5.24
NashMTL	0.60 (0.01)	4.61 (0.08)	154	1.68
GradDrop	0.60 (0.02)	4.60 (0.04)	141	1.15
PCGrad	0.60 (0.02)	4.75 (0.11)	114	2.89
CAGrad	0.61 (0.01)	4.69 (0.03)	138	3.64

Table 79: Results on the BPIC12C dataset for the CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.50 (0.01)	0.60 (0.01)	6.33 (0.05)	/	/
EW	0.77 (0.00)	0.61 (0.01)	6.40 (0.04)	172	-17.18
DWA	0.77 (0.00)	0.61 (0.02)	6.48 (0.04)	120	-16.91
RLW	0.77 (0.00)	0.61 (0.01)	6.44 (0.01)	191	-17.33
UW	0.78 (0.00)	0.60 (0.01)	6.83 (0.03)	70	-15.95
UW-O	0.78 (0.00)	0.60 (0.01)	6.86 (0.02)	60	-15.78
UW-SO	0.77 (0.00)	0.61 (0.01)	6.43 (0.06)	146	-17.29
GLS	0.78 (0.00)	0.60 (0.01)	6.86 (0.02)	55	-15.80
IMTL	0.78 (0.00)	0.60 (0.01)	6.78 (0.01)	80	-16.22
GradNorm	0.77 (0.00)	0.60 (0.01)	6.52 (0.05)	150	-17.54
NashMTL	0.77 (0.00)	0.61 (0.01)	6.47 (0.05)	138	-17.19
GradDrop	0.77 (0.00)	0.61 (0.01)	6.42 (0.03)	159	-16.85
PCGrad	0.77 (0.00)	0.60 (0.01)	6.49 (0.06)	137	-17.25
CAGrad	0.77 (0.00)	0.61 (0.01)	6.47 (0.07)	146	-17.20

Table 80: Results on the BPIC12C dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.50 (0.01)	0.60 (0.01)	/	/
EW	0.78 (0.00)	0.61 (0.01)	16	-27.92
DWA	0.78 (0.00)	0.60 (0.02)	19	-28.33
RLW	0.78 (0.00)	0.60 (0.02)	25	-28.18
UW	0.78 (0.00)	0.60 (0.01)	21	-28.19
UW-O	0.78 (0.00)	0.60 (0.01)	23	-28.13
UW-SO	0.78 (0.00)	0.60 (0.02)	18	-28.38
GLS	0.78 (0.01)	0.60 (0.01)	16	-27.62
IMTL	0.78 (0.00)	0.60 (0.02)	18	-28.51
GradNorm	0.78 (0.00)	0.60 (0.01)	21	-28.56
NashMTL	0.78 (0.00)	0.61 (0.02)	20	-27.84
GradDrop	0.78 (0.00)	0.60 (0.02)	20	-28.22
PCGrad	0.79 (0.00)	0.60 (0.01)	17	-28.26
CAGrad	0.78 (0.00)	0.60 (0.01)	24	-28.19

Table 81: Results on the BPIC12C dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.50 (0.01)	6.33 (0.05)	/	/
EW	0.77 (0.00)	6.42 (0.04)	167	-26.46
DWA	0.77 (0.00)	6.41 (0.02)	174	-26.43
RLW	0.77 (0.00)	6.44 (0.07)	159	-26.25
UW	0.78 (0.00)	6.83 (0.04)	68	-24.04
UW-O	0.78 (0.00)	6.86 (0.03)	61	-23.70
UW-SO	0.77 (0.00)	6.41 (0.05)	179	-26.58
GLS	0.78 (0.00)	6.86 (0.04)	57	-23.71
IMTL	0.77 (0.01)	6.74 (0.05)	95	-24.23
GradNorm	0.77 (0.01)	6.59 (0.33)	126	-25.10
NashMTL	0.77 (0.00)	6.49 (0.06)	134	-25.69
GradDrop	0.77 (0.00)	6.44 (0.07)	139	-26.03
PCGrad	0.77 (0.00)	6.43 (0.03)	169	-26.61
CAGrad	0.78 (0.00)	6.48 (0.06)	140	-26.28

Table 82: Results on the BPIC12C dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	0.60 (0.01)	6.33 (0.05)	/	/
EW	0.61 (0.01)	6.37 (0.06)	151	1.09
DWA	0.61 (0.01)	6.36 (0.03)	159	1.11
RLW	0.61 (0.01)	6.44 (0.02)	126	1.90
UW	0.60 (0.02)	6.54 (0.02)	128	1.76
UW-O	0.60 (0.01)	6.56 (0.02)	149	1.89
UW-SO	0.61 (0.01)	6.36 (0.09)	180	0.81
GLS	0.60 (0.01)	6.59 (0.07)	133	2.38
IMTL	0.60 (0.01)	6.50 (0.04)	144	1.09
GradNorm	0.59 (0.02)	6.51 (0.03)	139	0.85
NashMTL	0.61 (0.01)	6.46 (0.05)	131	2.18
GradDrop	0.61 (0.01)	6.36 (0.07)	143	1.20
PCGrad	0.60 (0.01)	6.30 (0.04)	181	0.07
CAGrad	0.60 (0.01)	6.37 (0.08)	150	0.45

Table 83: Results on the BPIC13I dataset for LSTM with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.52 (0.00)	1.40 (0.06)	9.65 (0.30)	/	/
EW	0.64 (0.01)	1.34 (0.06)	9.32 (0.31)	185	-10.69
DWA	0.64 (0.00)	1.33 (0.07)	9.22 (0.15)	198	-11.23
RLW	0.63 (0.01)	1.35 (0.05)	10.09 (0.48)	149	-7.33
UW	0.66 (0.00)	1.31 (0.07)	9.49 (0.11)	197	-12.16
UW-O	0.66 (0.01)	1.32 (0.06)	10.73 (0.68)	153	-7.52
UW-SO	0.64 (0.01)	1.32 (0.07)	9.28 (0.10)	197	-11.60
GLS	0.66 (0.00)	1.30 (0.07)	10.23 (0.55)	166	-10.02
IMTL	0.56 (0.06)	1.54 (0.05)	12.17 (0.26)	35	8.99
GradNorm	0.66 (0.00)	1.32 (0.07)	9.41 (0.17)	198	-12.25
NashMTL	0.64 (0.01)	1.33 (0.06)	9.32 (0.15)	195	-10.82
GradDrop	0.63 (0.01)	1.35 (0.06)	9.47 (0.37)	181	-9.63
PCGrad	0.66 (0.00)	1.32 (0.07)	9.65 (0.28)	195	-11.09
CAGrad	0.65 (0.01)	1.32 (0.07)	9.32 (0.23)	197	-11.71

Table 84: Results on the BPIC13I dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.52 (0.00)	1.40 (0.06)	/	/
EW	0.66 (0.01)	1.28 (0.02)	114	-18.29
DWA	0.66 (0.01)	1.28 (0.02)	82	-18.23
RLW	0.66 (0.00)	1.28 (0.02)	108	-18.24
UW	0.66 (0.01)	1.28 (0.02)	84	-18.17
UW-O	0.66 (0.00)	1.29 (0.02)	80	-18.05
UW-SO	0.66 (0.01)	1.28 (0.01)	112	-18.52
GLS	0.66 (0.01)	1.28 (0.02)	122	-18.81
IMTL	0.66 (0.00)	1.30 (0.02)	86	-17.97
GradNorm	0.66 (0.01)	1.29 (0.01)	94	-18.12
NashMTL	0.66 (0.01)	1.28 (0.02)	90	-17.90
GradDrop	0.66 (0.01)	1.27 (0.02)	118	-18.69
PCGrad	0.66 (0.00)	1.29 (0.01)	92	-18.04
CAGrad	0.66 (0.01)	1.28 (0.01)	96	-18.28

Table 85: Results on the BPIC13I dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.52 (0.00)	9.65 (0.30)	/	/
EW	0.64 (0.00)	9.13 (0.42)	188	-14.34
DWA	0.64 (0.01)	9.05 (0.25)	196	-14.82
RLW	0.63 (0.00)	9.47 (0.02)	179	-12.29
UW	0.66 (0.00)	9.34 (0.25)	184	-15.77
UW-O	0.66 (0.00)	9.59 (0.01)	199	-14.56
UW-SO	0.63 (0.00)	9.05 (0.14)	194	-14.71
GLS	0.66 (0.00)	9.39 (0.14)	200	-15.37
IMTL	0.66 (0.00)	9.82 (0.65)	159	-13.08
GradNorm	0.66 (0.00)	9.25 (0.19)	200	-15.99
NashMTL	0.63 (0.00)	9.39 (0.15)	177	-12.89
GradDrop	0.62 (0.02)	9.28 (0.24)	197	-12.42
PCGrad	0.66 (0.01)	9.31 (0.14)	194	-15.49
CAGrad	0.64 (0.00)	9.16 (0.21)	188	-14.86

Table 86: Results on the BPIC13I dataset for LSTM with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	1.40 (0.06)	9.65 (0.30)	/	/
EW	1.50 (0.00)	9.88 (0.70)	159	4.54
DWA	1.46 (0.04)	9.50 (0.19)	200	1.33
RLW	1.45 (0.06)	9.76 (0.01)	197	2.10
UW	1.35 (0.04)	9.63 (0.11)	198	-1.80
UW-O	1.29 (0.00)	9.81 (0.03)	197	-3.31
UW-SO	1.39 (0.01)	9.31 (0.28)	198	-2.10
GLS	1.33 (0.03)	10.14 (0.32)	156	-0.18
IMTL	1.32 (0.00)	9.16 (0.30)	198	-5.35
GradNorm	1.32 (0.01)	9.29 (0.18)	197	-4.66
NashMTL	1.40 (0.08)	9.42 (0.28)	200	-1.39
GradDrop	1.46 (0.07)	9.53 (0.08)	199	1.31
PCGrad	1.40 (0.10)	9.51 (0.01)	189	-0.83
CAGrad	1.34 (0.02)	9.18 (0.31)	200	-4.73

Table 87: Results on the BPIC13I dataset for the CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.46 (0.01)	2.23 (0.61)	15.92 (0.33)	/	/
EW	0.52 (0.00)	1.57 (0.06)	12.08 (0.34)	125	-21.87
DWA	0.52 (0.01)	1.56 (0.06)	12.07 (0.34)	163	-22.56
RLW	0.52 (0.00)	1.56 (0.06)	12.06 (0.26)	192	-21.97
UW	0.54 (0.01)	1.55 (0.06)	12.07 (0.31)	155	-23.75
UW-O	0.54 (0.01)	1.55 (0.07)	11.99 (0.28)	112	-23.93
UW-SO	0.53 (0.00)	1.56 (0.06)	12.01 (0.38)	159	-22.98
GLS	0.54 (0.01)	1.55 (0.07)	12.01 (0.32)	126	-23.85
IMTL	0.50 (0.04)	1.57 (0.08)	12.10 (0.44)	105	-20.90
GradNorm	0.51 (0.01)	1.57 (0.07)	12.26 (0.31)	119	-20.97
NashMTL	0.46 (0.00)	1.58 (0.07)	12.29 (0.29)	59	-16.98
GradDrop	0.51 (0.00)	1.57 (0.06)	12.12 (0.18)	160	-21.58
PCGrad	0.52 (0.01)	1.56 (0.06)	12.10 (0.26)	148	-22.05
CAGrad	0.53 (0.00)	1.55 (0.06)	12.03 (0.31)	156	-23.05

Table 88: Results on the BPIC13I dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.46 (0.01)	2.23 (0.61)	/	/
EW	0.53 (0.01)	1.52 (0.00)	72	-23.45
DWA	0.54 (0.01)	1.52 (0.00)	91	-24.08
RLW	0.53 (0.01)	1.52 (0.00)	100	-23.64
UW	0.53 (0.00)	1.52 (0.00)	92	-23.30
UW-O	0.53 (0.01)	1.52 (0.00)	66	-23.67
UW-SO	0.53 (0.01)	1.52 (0.00)	86	-23.94
GLS	0.53 (0.01)	1.52 (0.00)	96	-24.03
IMTL	0.53 (0.00)	1.51 (0.00)	93	-23.72
GradNorm	0.53 (0.00)	1.52 (0.01)	136	-23.94
NashMTL	0.53 (0.01)	1.52 (0.00)	76	-23.76
GradDrop	0.53 (0.01)	1.52 (0.00)	92	-23.92
PCGrad	0.53 (0.01)	1.51 (0.00)	94	-23.82
CAGrad	0.54 (0.01)	1.51 (0.00)	98	-24.58

Table 89: Results on the BPIC13I dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.46 (0.01)	15.92 (0.33)	/	/
EW	0.48 (0.04)	13.73 (2.53)	102	-9.18
DWA	0.51 (0.01)	11.92 (0.03)	164	-18.22
RLW	0.45 (0.00)	13.92 (2.27)	92	-5.53
UW	0.47 (0.07)	13.81 (2.48)	74	-8.03
UW-O	0.54 (0.01)	11.80 (0.21)	102	-21.19
UW-SO	0.52 (0.00)	11.80 (0.04)	168	-19.77
GLS	0.53 (0.01)	11.87 (0.26)	91	-20.49
IMTL	0.52 (0.01)	11.80 (0.25)	98	-19.73
GradNorm	0.53 (0.00)	12.09 (0.15)	116	-19.72
NashMTL	0.49 (0.05)	11.93 (0.51)	78	-15.40
GradDrop	0.51 (0.00)	11.95 (0.03)	164	-17.49
PCGrad	0.52 (0.00)	11.82 (0.15)	179	-19.54
CAGrad	0.53 (0.00)	11.86 (0.10)	190	-20.32

Table 90: Results on the BPIC13I dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	2.23 (0.61)	15.92 (0.33)	/	/
EW	2.50 (0.01)	15.75 (0.19)	9	5.57
DWA	2.50 (0.00)	15.76 (0.18)	5	5.55
RLW	2.50 (0.01)	15.74 (0.15)	10	5.37
UW	2.52 (0.01)	15.76 (0.20)	8	5.84
UW-O	1.54 (0.01)	12.18 (0.22)	66	-27.30
UW-SO	2.50 (0.01)	15.74 (0.18)	6	5.56
GLS	2.50 (0.01)	15.75 (0.18)	6	5.55
GradNorm	2.49 (0.02)	15.76 (0.18)	8	5.37
NashMTL	1.54 (0.01)	12.17 (0.23)	36	-27.30
GradDrop	2.51 (0.01)	15.75 (0.18)	4	5.78
PCGrad	2.51 (0.01)	15.76 (0.19)	6	5.63
CAGrad	2.51 (0.01)	15.74 (0.17)	8	5.71

Table 91: Results on the BPIC13I dataset for Transformer with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.59 (0.05)	1.23 (0.08)	6.77 (0.60)	/	/
EW	0.68 (0.01)	1.25 (0.04)	6.19 (0.45)	54	-7.01
DWA	0.67 (0.01)	1.35 (0.12)	6.28 (0.21)	47	-3.43
RLW	0.67 (0.01)	1.24 (0.04)	6.50 (0.21)	52	-5.69
UW	0.70 (0.01)	1.30 (0.09)	6.74 (0.30)	45	-4.08
GLS	0.71 (0.01)	1.26 (0.07)	7.57 (0.34)	39	-2.10
IMTL	0.71 (0.01)	1.33 (0.11)	7.32 (0.17)	40	-1.01
GradNorm	0.70 (0.00)	1.29 (0.08)	7.08 (0.32)	39	-2.62
NashMTL	0.68 (0.01)	1.29 (0.05)	6.47 (0.16)	38	-4.81
GradDrop	0.66 (0.01)	1.29 (0.08)	6.24 (0.33)	43	-4.96
PCGrad	0.68 (0.01)	1.32 (0.12)	6.25 (0.36)	55	-5.22
CAGrad	0.69 (0.01)	1.38 (0.22)	6.22 (0.29)	48	-4.00

Table 92: Results on the BPIC15-1 dataset for LSTM with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.17 (0.01)	2.03 (0.20)	/	/
EW	0.44 (0.00)	2.01 (0.21)	19	-83.74
DWA	0.44 (0.01)	2.00 (0.21)	20	-84.09
RLW	0.44 (0.00)	1.99 (0.20)	24	-83.62
UW	0.45 (0.00)	1.99 (0.21)	21	-84.90
UW-O	0.44 (0.01)	2.06 (0.23)	153	-80.06
UW-SO	0.44 (0.00)	2.00 (0.21)	19	-83.78
GLS	0.45 (0.01)	2.01 (0.22)	24	-86.00
IMTL	0.44 (0.01)	1.95 (0.19)	29	-83.76
GradNorm	0.44 (0.00)	1.96 (0.21)	27	-83.38
NashMTL	0.45 (0.01)	1.99 (0.20)	24	-86.60
GradDrop	0.45 (0.00)	1.99 (0.21)	22	-85.83
PCGrad	0.44 (0.00)	1.99 (0.22)	18	-84.23
CAGrad	0.45 (0.00)	1.94 (0.21)	29	-85.79

Table 93: Results on the BPIC15-1 dataset for LSTM with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.17 (0.01)	13.65 (0.44)	/	/
EW	0.30 (0.00)	14.92 (0.92)	91	-36.10
DWA	0.30 (0.00)	14.18 (0.31)	127	-37.66
RLW	0.30 (0.00)	14.98 (0.81)	109	-35.01
UW	0.40 (0.00)	16.40 (0.91)	93	-58.71
UW-O	0.41 (0.01)	17.99 (1.26)	89	-55.74
UW-SO	0.25 (0.01)	16.54 (0.95)	198	-14.16
GLS	0.40 (0.00)	17.53 (0.67)	81	-56.61
IMTL	0.42 (0.01)	21.63 (1.47)	196	-46.98
GradNorm	0.40 (0.00)	19.84 (1.35)	198	-48.53
NashMTL	0.25 (0.00)	17.55 (1.24)	197	-10.99
GradDrop	0.22 (0.00)	17.86 (1.04)	186	-1.69
CAGrad	0.36 (0.01)	14.96 (0.50)	112	-54.05

Table 94: Results on the BPIC15-1 dataset for CNN with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.11 (0.00)	2.08 (0.22)	35.15 (13.38)	/	/
EW	0.20 (0.01)	2.08 (0.22)	27.58 (1.43)	167	-33.70
DWA	0.19 (0.02)	2.08 (0.22)	30.00 (6.43)	138	-28.47
RLW	0.19 (0.02)	2.08 (0.22)	32.13 (3.24)	123	-28.17
UW	0.28 (0.01)	2.09 (0.23)	32.04 (0.93)	157	-53.68
UW-O	0.28 (0.00)	2.09 (0.23)	32.29 (1.88)	132	-55.65
UW-SO	0.20 (0.01)	2.08 (0.22)	27.04 (1.21)	188	-35.70
GLS	0.22 (0.08)	2.09 (0.23)	36.29 (4.34)	83	-33.38
IMTL	0.25 (0.01)	2.08 (0.22)	37.73 (0.75)	52	-39.89
GradNorm	0.21 (0.06)	2.10 (0.21)	33.66 (7.88)	99	-32.89
NashMTL	0.21 (0.01)	2.09 (0.22)	28.57 (1.24)	135	-36.49
GradDrop	0.20 (0.00)	2.08 (0.22)	27.85 (1.01)	190	-33.76
PCGrad	0.24 (0.01)	2.08 (0.23)	27.28 (0.33)	196	-48.87
CAGrad	0.21 (0.01)	2.08 (0.22)	28.45 (2.44)	195	-37.29

Table 95: Results on the BPIC15-1 dataset for CNN with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.11 (0.00)	2.08 (0.22)	/	/
EW	0.26 (0.03)	2.08 (0.22)	21	-69.86
DWA	0.28 (0.01)	2.08 (0.23)	26	-76.36
RLW	0.27 (0.01)	2.11 (0.20)	27	-72.50
UW	0.27 (0.03)	2.09 (0.23)	24	-74.01
UW-O	0.29 (0.01)	2.08 (0.22)	58	-81.05
UW-SO	0.28 (0.02)	2.08 (0.22)	26	-76.77
GLS	0.26 (0.00)	2.09 (0.22)	40	-70.24
IMTL	0.24 (0.01)	2.08 (0.22)	34	-61.89
GradNorm	0.24 (0.02)	2.08 (0.23)	33	-59.56
NashMTL	0.27 (0.01)	2.09 (0.22)	12	-72.17
GradDrop	0.28 (0.01)	2.09 (0.22)	21	-76.92
PCGrad	0.27 (0.01)	2.08 (0.22)	26	-75.54
CAGrad	0.27 (0.02)	2.08 (0.22)	28	-75.91

Table 96: Results on the BPIC15-1 dataset for CNN with task combination NAP+RTP

MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.11 (0.00)	35.15 (13.38)	/	/
EW	0.20 (0.01)	26.67 (1.22)	196	-53.67
DWA	0.20 (0.00)	26.68 (1.30)	198	-54.42
RLW	0.20 (0.01)	27.82 (1.44)	199	-51.26
UW	0.27 (0.02)	33.38 (1.90)	141	-74.74
UW-O	0.27 (0.01)	33.48 (0.93)	68	-76.63
GLS	0.28 (0.01)	33.42 (2.60)	116	-81.49
CAGrad	0.25 (0.01)	27.88 (1.76)	183	-73.92

Table 97: Results on the BPIC15-1 dataset for CNN with task combination NTP+RTP

MTO	NTP	RTP	Best Epoch	Δ_m
STL	2.08 (0.22)	35.15 (13.38)	/	/
EW	2.08 (0.22)	26.70 (0.21)	199	-11.87
DWA	2.08 (0.22)	26.87 (1.06)	199	-11.59
RLW	2.40 (0.50)	36.07 (13.58)	137	9.18
UW	2.08 (0.22)	27.15 (0.45)	199	-11.28
GLS	2.08 (0.22)	28.83 (0.36)	185	-8.97
GradNorm	2.08 (0.23)	36.06 (0.79)	84	1.37
NashMTL	2.14 (0.28)	29.68 (0.99)	118	-6.18
GradDrop	2.40 (0.50)	34.95 (14.48)	137	7.58
PCGrad	2.08 (0.22)	27.27 (0.44)	196	-11.20

Table 98: Results on the BPIC15-1 dataset for Transformer with task combination NAP+NTP

MTO	NAP	NTP	Best Epoch	Δ_m
STL	0.14 (0.01)	2.03 (0.22)	/	/
EW	0.51 (0.01)	1.99 (0.19)	29	-131.11
DWA	0.51 (0.01)	1.98 (0.19)	29	-131.91
RLW	0.51 (0.01)	1.95 (0.19)	33	-131.33
UW	0.51 (0.01)	1.97 (0.20)	31	-131.04
UW-O	0.50 (0.01)	2.02 (0.21)	45	-127.79
UW-SO	0.51 (0.00)	1.95 (0.18)	32	-131.52
GLS	0.51 (0.01)	1.99 (0.20)	35	-130.01
IMTL	0.51 (0.01)	1.98 (0.17)	28	-130.37
GradNorm	0.51 (0.00)	1.94 (0.16)	33	-130.18
NashMTL	0.51 (0.00)	1.99 (0.17)	30	-131.49
GradDrop	0.50 (0.00)	1.96 (0.22)	34	-129.38
PCGrad	0.51 (0.01)	1.96 (0.19)	33	-132.25
CAGrad	0.51 (0.01)	1.95 (0.22)	29	-132.24

Table 99: Results on the BPIC15-1 dataset for Transformer with task combination NAP+RTP

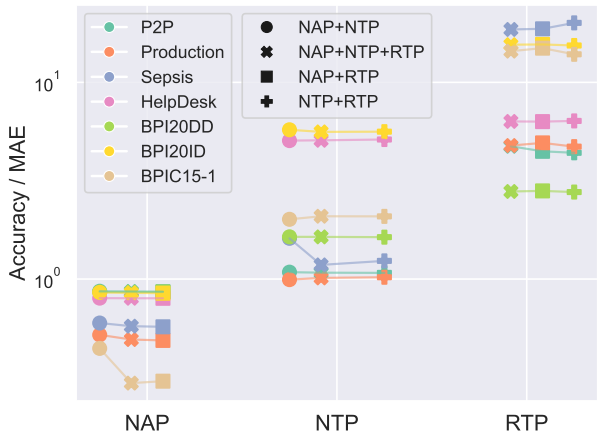
MTO	NAP	RTP	Best Epoch	Δ_m
STL	0.14 (0.01)	20.84 (2.22)	/	/
EW	0.17 (0.02)	19.73 (2.41)	57	-13.30
DWA	0.17 (0.03)	20.05 (2.25)	53	-13.12
RLW	0.16 (0.04)	21.57 (2.78)	48	-5.07
UW	0.37 (0.06)	20.64 (0.71)	60	-81.28
UW-O	0.50 (0.01)	24.11 (0.72)	54	-117.04
UW-SO	0.17 (0.03)	20.64 (1.74)	41	-10.23
GLS	0.47 (0.02)	25.15 (0.83)	40	-106.82
IMTL	0.50 (0.01)	27.35 (0.31)	54	-111.11
GradNorm	0.44 (0.06)	25.27 (0.77)	31	-94.28
NashMTL	0.18 (0.01)	21.80 (0.33)	31	-10.83
GradDrop	0.15 (0.01)	20.03 (0.66)	49	-3.49
PCGrad	0.26 (0.01)	21.26 (0.66)	43	-40.21
CAGrad	0.44 (0.01)	21.57 (0.36)	59	-103.13

Table 100: Results on the BPIC15-1 dataset for Transformer with task combination NTP+RTP

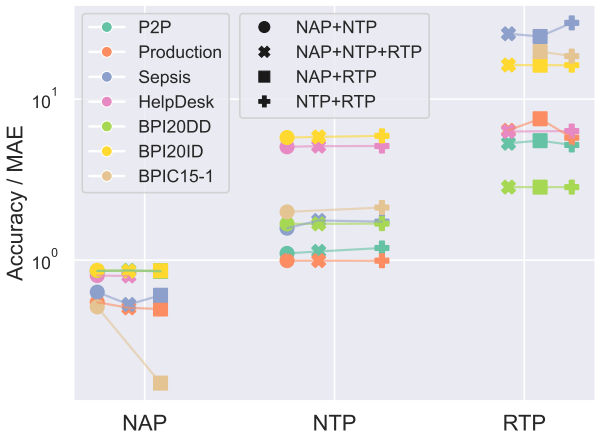
MTO	NTP	RTP	Best Epoch	Δ_m
STL	2.03 (0.22)	20.84 (2.22)	/	/
EW	2.12 (0.22)	18.50 (1.05)	73	-3.41
DWA	2.12 (0.22)	19.40 (1.14)	59	-1.26
RLW	2.12 (0.22)	21.04 (1.55)	53	2.65
UW	2.11 (0.23)	29.46 (20.10)	47	22.49
UW-O	2.16 (0.20)	19.28 (1.59)	66	-0.72
UW-SO	2.12 (0.22)	18.62 (1.92)	72	-3.13
GLS	2.09 (0.21)	20.43 (1.25)	58	0.50
IMTL	2.03 (0.25)	21.37 (1.75)	56	1.10
GradNorm	2.06 (0.24)	20.47 (0.84)	56	-0.21
NashMTL	2.12 (0.23)	18.71 (3.58)	69	-2.99
GradDrop	2.12 (0.22)	19.20 (0.79)	50	-1.74
PCGrad	2.11 (0.24)	18.56 (0.98)	62	-3.65
CAGrad	2.08 (0.22)	19.66 (2.06)	52	-1.63

Impact of Task Combinations

In Figures 1a and 1b, we complement the main paper by showing the task performances for different task combinations for LSTM and Transformer.



(a) LSTM



(b) Transformer

Figure 1: Impact of task combinations on performance. Results are based on EW with the (a) LSTM and (b) Transformer model.

Ablation: Simplified Model

In the following, we report the full results for our ablation study on the simplified CNN model.

Table 101: Results on the P2P dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.33 (0.04)	1.15 (0.09)	6.04 (0.31)	/	/
EW	0.85 (0.01)	1.17 (0.07)	6.07 (0.32)	55	-52.63
DWA	0.85 (0.01)	1.16 (0.08)	6.09 (0.32)	54	-52.79
RLW	0.84 (0.02)	1.17 (0.10)	6.28 (0.30)	54	-51.36
UW	0.83 (0.02)	1.17 (0.09)	6.56 (0.19)	28	-48.51
UW-O	0.83 (0.01)	1.21 (0.10)	6.67 (0.20)	14	-46.99
UW-SO	0.84 (0.01)	1.17 (0.07)	6.11 (0.12)	58	-52.25
GLS	0.84 (0.02)	1.20 (0.10)	6.66 (0.24)	19	-48.37
IMTL	0.85 (0.01)	1.18 (0.07)	6.61 (0.30)	22	-49.20
GradNorm	0.84 (0.01)	1.16 (0.06)	6.20 (0.26)	58	-51.89
NashMTL	0.84 (0.01)	1.18 (0.07)	6.13 (0.26)	49	-51.89
GradDrop	0.85 (0.01)	1.17 (0.07)	6.09 (0.28)	55	-52.38
PCGrad	0.84 (0.02)	1.19 (0.10)	6.17 (0.48)	55	-51.07
CAGrad	0.85 (0.01)	1.18 (0.08)	6.37 (0.36)	36	-50.72

Table 102: Results on the Production dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.31 (0.00)	1.01 (0.16)	16.10 (1.85)	/	/
EW	0.38 (0.04)	1.16 (0.20)	16.56 (2.46)	21	-1.58
DWA	0.38 (0.03)	1.16 (0.18)	17.20 (1.60)	14	0.15
RLW	0.40 (0.02)	1.19 (0.09)	17.60 (0.95)	19	-0.30
UW	0.43 (0.02)	1.16 (0.20)	15.51 (2.04)	46	-8.69
UW-O	0.43 (0.01)	1.06 (0.14)	13.42 (0.20)	28	-16.80
UW-SO	0.39 (0.05)	1.09 (0.20)	14.21 (3.39)	57	-10.33
GLS	0.43 (0.03)	1.07 (0.16)	14.41 (0.05)	39	-15.00
IMTL	0.43 (0.03)	1.06 (0.15)	13.41 (0.72)	39	-17.10
GradNorm	0.40 (0.03)	1.16 (0.26)	16.15 (2.89)	36	-5.03
NashMTL	0.41 (0.01)	1.04 (0.16)	11.16 (0.38)	70	-19.59
GradDrop	0.41 (0.02)	1.19 (0.19)	16.48 (1.35)	19	-4.06
PCGrad	0.43 (0.00)	1.16 (0.06)	16.27 (1.70)	12	-7.15
CAGrad	0.43 (0.03)	1.10 (0.12)	14.23 (2.20)	50	-14.19

Table 103: Results on the Helpdesk dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.65 (0.01)	6.13 (0.06)	7.40 (0.11)	/	/
EW	0.76 (0.01)	6.09 (0.10)	7.63 (0.08)	45	-4.39
DWA	0.76 (0.01)	6.09 (0.09)	7.70 (0.09)	31	-4.23
RLW	0.76 (0.01)	6.13 (0.06)	7.71 (0.20)	49	-4.00
UW	0.77 (0.00)	6.09 (0.13)	7.75 (0.07)	45	-4.42
UW-O	0.77 (0.00)	6.03 (0.08)	7.66 (0.11)	44	-5.37
UW-SO	0.76 (0.00)	6.09 (0.09)	7.61 (0.14)	52	-4.75
GLS	0.77 (0.01)	6.03 (0.06)	7.69 (0.06)	47	-5.11
IMTL	0.77 (0.00)	6.00 (0.05)	7.52 (0.03)	53	-6.06
GradNorm	0.77 (0.00)	6.09 (0.07)	7.73 (0.07)	47	-4.68
NashMTL	0.76 (0.01)	6.05 (0.11)	7.60 (0.15)	38	-4.97
GradDrop	0.76 (0.01)	6.11 (0.17)	7.65 (0.24)	44	-4.30
PCGrad	0.76 (0.01)	6.05 (0.08)	7.62 (0.11)	47	-4.97
CAGrad	0.76 (0.01)	6.06 (0.07)	7.62 (0.10)	45	-4.93

Table 104: Results on the Sepsis dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.43 (0.01)	1.96 (0.14)	26.58 (1.12)	/	/
EW	0.51 (0.00)	2.05 (0.13)	28.56 (0.58)	28	-2.21
DWA	0.51 (0.01)	2.04 (0.14)	28.74 (0.61)	21	-2.47
RLW	0.49 (0.02)	2.04 (0.13)	28.03 (1.18)	47	-1.56
UW	0.55 (0.01)	2.03 (0.15)	28.61 (0.63)	36	-5.65
UW-O	0.55 (0.01)	2.01 (0.14)	29.06 (0.57)	23	-5.32
UW-SO	0.51 (0.02)	2.00 (0.12)	26.88 (1.27)	83	-4.99
GLS	0.55 (0.01)	2.01 (0.13)	29.21 (0.77)	18	-5.37
IMTL	0.56 (0.01)	1.94 (0.13)	28.69 (0.77)	30	-7.58
GradNorm	0.55 (0.01)	1.94 (0.12)	27.67 (0.79)	86	-8.41
NashMTL	0.49 (0.00)	2.03 (0.12)	27.47 (1.70)	56	-2.46
GradDrop	0.49 (0.04)	2.04 (0.18)	28.76 (0.99)	53	-0.96
PCGrad	0.52 (0.01)	2.00 (0.15)	27.76 (1.44)	58	-5.19
CAGrad	0.52 (0.01)	1.97 (0.11)	27.04 (1.28)	74	-6.41

Table 105: Results on the BPIC12C dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.46 (0.00)	0.68 (0.01)	7.29 (0.04)	/	/
EW	0.74 (0.01)	0.71 (0.01)	7.29 (0.03)	26	-18.75
DWA	0.74 (0.00)	0.70 (0.00)	7.32 (0.03)	32	-19.11
RLW	0.74 (0.01)	0.71 (0.01)	7.29 (0.06)	31	-18.41
UW	0.76 (0.00)	0.69 (0.01)	7.33 (0.04)	37	-20.69
UW-O	0.76 (0.00)	0.70 (0.01)	7.32 (0.03)	31	-20.70
UW-SO	0.75 (0.00)	0.70 (0.00)	7.25 (0.04)	44	-19.73
GLS	0.76 (0.00)	0.69 (0.01)	7.34 (0.02)	27	-20.50
IMTL	0.76 (0.00)	0.69 (0.01)	7.30 (0.06)	29	-20.64
GradNorm	0.76 (0.00)	0.69 (0.01)	7.34 (0.02)	27	-20.60
NashMTL	0.74 (0.01)	0.69 (0.01)	7.24 (0.04)	30	-19.78
GradDrop	0.74 (0.00)	0.70 (0.00)	7.28 (0.03)	33	-19.06
PCGrad	0.74 (0.00)	0.70 (0.02)	7.30 (0.03)	25	-19.26
CAGrad	0.75 (0.01)	0.69 (0.01)	7.34 (0.04)	25	-19.87

Table 106: Results on the BPIC13I dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.45 (0.00)	2.58 (0.07)	15.97 (0.35)	/	/
EW	0.39 (0.02)	2.57 (0.08)	15.93 (0.37)	7	4.19
DWA	0.40 (0.07)	1.97 (0.49)	15.09 (1.61)	25	-5.65
UW	0.41 (0.04)	2.80 (1.42)	14.75 (2.37)	19	3.70
UW-O	0.45 (0.00)	1.58 (0.07)	12.30 (0.28)	21	-20.65
UW-SO	0.41 (0.01)	1.62 (0.07)	12.61 (0.63)	51	-16.05
GLS	0.45 (0.00)	1.58 (0.07)	12.31 (0.27)	21	-20.66
GradNorm	0.38 (0.04)	2.24 (0.61)	14.76 (2.32)	24	-1.64
NashMTL	0.43 (0.02)	1.58 (0.07)	12.28 (0.28)	22	-18.66
GradDrop	0.40 (0.04)	2.88 (2.26)	15.18 (2.26)	30	5.84
PCGrad	0.39 (0.06)	2.33 (0.72)	14.75 (2.38)	22	-0.70
CAGrad	0.44 (0.02)	1.63 (0.08)	12.64 (0.37)	53	-18.15

Table 107: Results on the BPIC15-1 dataset for the simplified CNN model with task combination NAP+NTP+RTP

MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.12 (0.00)	3.06 (0.27)	52.60 (0.69)	/	/
EW	0.14 (0.01)	3.05 (0.22)	52.55 (0.72)	2	-6.92
DWA	0.15 (0.03)	3.04 (0.23)	52.34 (1.05)	6	-9.38
UW	0.16 (0.01)	3.07 (0.21)	52.53 (0.77)	4	-11.63
UW-O	0.18 (0.02)	2.13 (0.22)	39.39 (0.75)	9	-35.09
UW-SO	0.14 (0.01)	3.09 (0.19)	52.55 (0.73)	2	-4.99
GLS	0.21 (0.00)	2.12 (0.21)	39.76 (0.69)	20	-44.02
GradNorm	0.15 (0.04)	3.05 (0.22)	52.56 (0.71)	6	-9.91
NashMTL	0.16 (0.01)	2.11 (0.23)	39.46 (0.46)	11	-30.69
GradDrop	0.16 (0.03)	3.63 (0.94)	52.58 (0.69)	9	-5.03
PCGrad	0.13 (0.03)	3.08 (0.20)	52.58 (0.69)	1	-2.82
CAGrad	0.14 (0.02)	2.74 (0.69)	47.84 (8.72)	15	-12.87

Table 108: Results on the BPIC20DD dataset for the simplified CNN model with task combination NAP+NTP+RTP

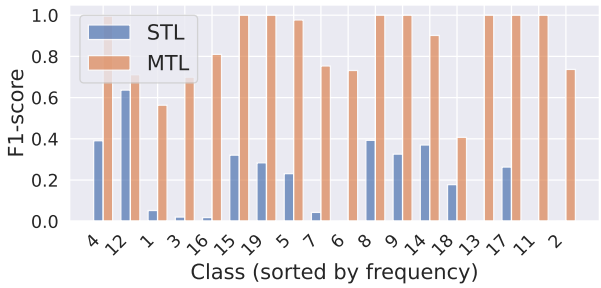
MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.66 (0.00)	1.71 (0.03)	2.85 (0.02)	/	/
EW	0.86 (0.01)	1.72 (0.03)	2.86 (0.01)	43	-9.83
DWA	0.86 (0.01)	1.72 (0.02)	2.86 (0.02)	40	-9.80
RLW	0.86 (0.01)	1.72 (0.03)	2.86 (0.01)	48	-9.73
UW	0.86 (0.01)	1.74 (0.01)	2.86 (0.01)	30	-9.40
UW-O	0.86 (0.01)	1.73 (0.02)	2.87 (0.02)	39	-9.54
UW-SO	0.86 (0.01)	1.72 (0.03)	2.86 (0.02)	41	-9.83
GLS	0.86 (0.01)	1.73 (0.03)	2.86 (0.02)	42	-9.68
IMTL	0.86 (0.01)	1.74 (0.03)	2.87 (0.02)	32	-9.43
GradNorm	0.86 (0.01)	1.72 (0.03)	2.86 (0.02)	40	-9.82
NashMTL	0.86 (0.01)	1.72 (0.03)	2.86 (0.01)	37	-9.79
GradDrop	0.86 (0.01)	1.73 (0.02)	2.87 (0.02)	35	-9.54
PCGrad	0.86 (0.01)	1.72 (0.02)	2.86 (0.02)	39	-9.78
CAGrad	0.86 (0.00)	1.73 (0.02)	2.86 (0.01)	36	-9.71

Table 109: Results on the BPIC20ID dataset for the simplified CNN model with task combination NAP+NTP+RTP

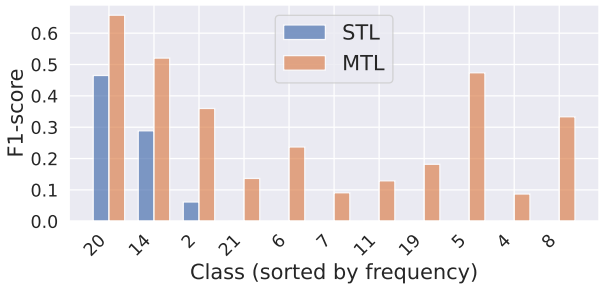
MTO	NAP	NTP	RTP	Best Epoch	Δ_m
STL	0.47 (0.01)	6.87 (0.08)	18.81 (0.26)	/	/
EW	0.78 (0.01)	6.80 (0.03)	18.79 (0.13)	35	-21.86
DWA	0.78 (0.01)	6.76 (0.06)	18.83 (0.23)	26	-21.94
RLW	0.76 (0.01)	6.82 (0.13)	19.02 (0.22)	31	-20.42
UW	0.81 (0.00)	6.82 (0.05)	19.00 (0.10)	42	-23.79
UW-O	0.82 (0.00)	6.85 (0.02)	19.41 (0.28)	47	-23.18
UW-SO	0.81 (0.01)	6.75 (0.05)	18.77 (0.20)	48	-24.10
GLS	0.81 (0.01)	6.88 (0.01)	19.39 (0.23)	40	-22.89
IMTL	0.81 (0.01)	6.77 (0.12)	19.17 (0.05)	31	-23.29
GradNorm	0.80 (0.01)	6.79 (0.13)	18.78 (0.31)	40	-23.38
NashMTL	0.78 (0.00)	6.90 (0.09)	18.97 (0.14)	27	-21.28
GradDrop	0.78 (0.00)	6.85 (0.03)	18.74 (0.10)	33	-21.69
PCGrad	0.79 (0.01)	6.78 (0.12)	18.67 (0.12)	42	-23.23
CAGrad	0.79 (0.01)	6.76 (0.02)	18.63 (0.18)	45	-23.55

Impact of MTL on Next Activity Prediction: The Case of Infrequent Classes

In this section, we investigate the extent to which jointly learning multiple PPM tasks through MTL improves the accuracy of next activity prediction (NAP), with a particular focus on infrequent activities within the process. For conciseness, we report results for the ”NAP+NTP+RTP” task combination. Similar performance trends were observed across other task combinations. These findings suggest that MTL effectively leverages auxiliary signals from time-based tasks (NTP and RTP), encouraging the model to learn richer and more structured latent representations, which in turn improves NAP accuracy—particularly for infrequent classes.

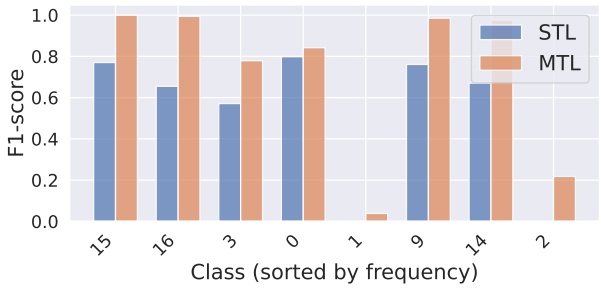


(a) P2P event log



(b) Production Event Log

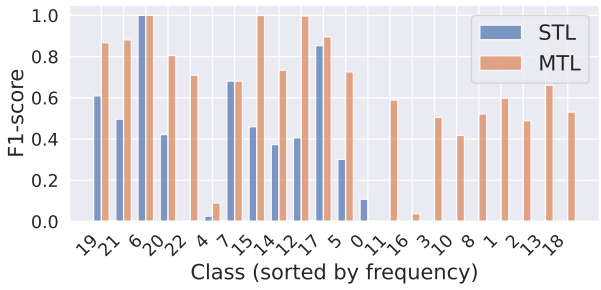
Figure 2: Per-class F1-score comparison between MTL and STL models using CNN architecture. The MTL model employs the NAP+NTP+RTP task combination. Activity classes are sorted by frequency from left to right.



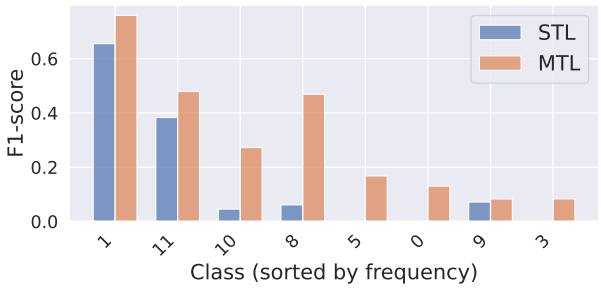
(a) BPIC20DD (Domestic Declarations)

(b) BPIC20ID (International Declarations)

Figure 3: Per-class F1-score comparison between MTL and STL models using CNN architecture. The MTL model employs the NAP+NTP+RTP task combination. Activity classes are sorted by frequency from left to right.



(a) BPIC12C

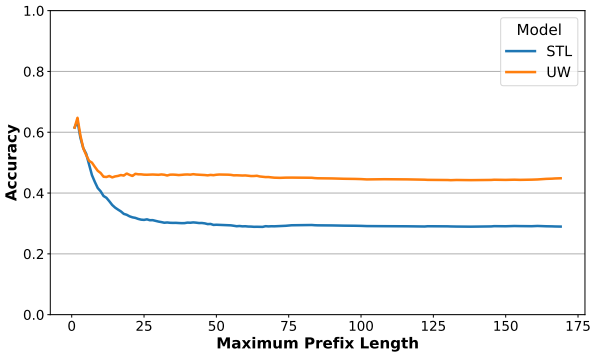


(b) BPIC13I

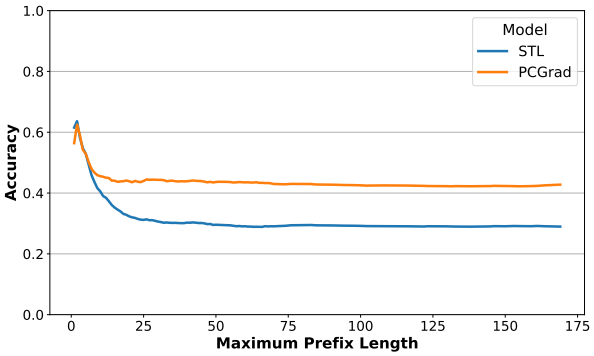
Figure 4: Per-class F1-score comparison between MTL and STL models using CNN architecture. The MTL model employs the NAP+NTP+RTP task combination. Activity classes are sorted by frequency from left to right.

Impact of MTL on Next Activity Prediction: Prefix Length Analysis

In this section, we examine how MTL affects the accuracy of next activity prediction (NAP) across event prefixes of varying lengths. For clarity and brevity, we focus on the ”NAP+NTP+RTP” task combination and present results for two representative MTO methods: UW and PCGrad. Consistent trends were observed across other task combinations, and MTO methods. Notably, both MTO methods consistently outperform the STL baseline across the full spectrum of prefix lengths, indicating the robustness of MTL in leveraging temporal context at different stages of a process.

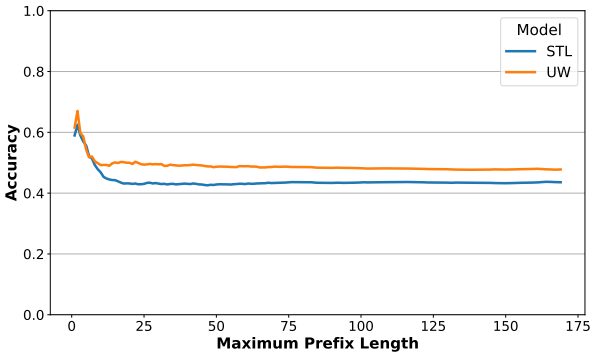


(a) UW

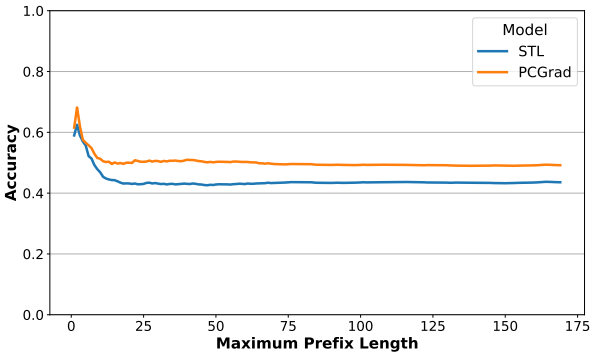


(b) PCGrad

Figure 5: Impact of MTL on NAP accuracy in Production log (CNN model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

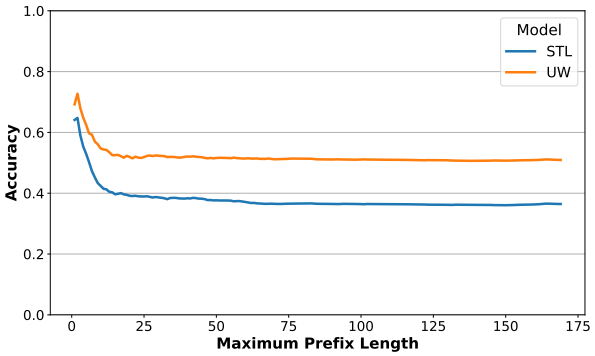


(a) UW

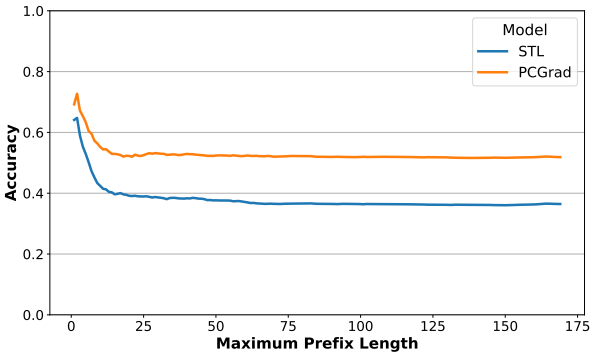


(b) PCGrad

Figure 6: Impact of MTL on NAP accuracy in Production log (LSTM model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

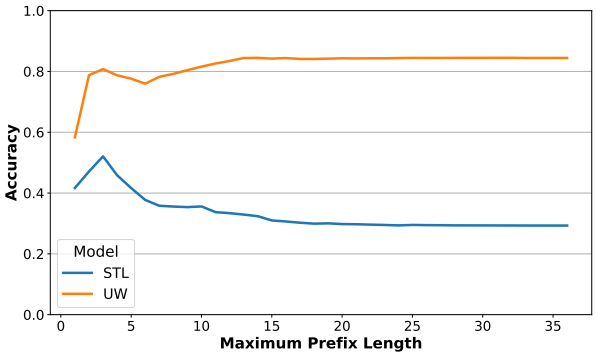


(a) UW

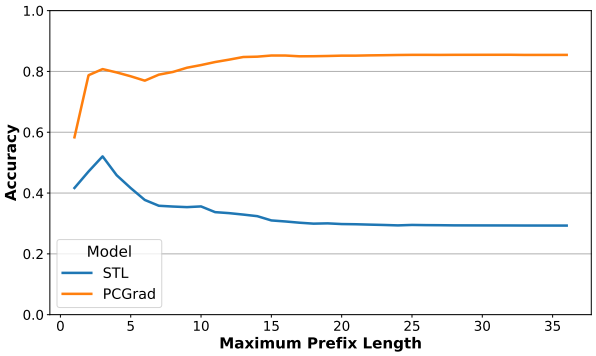


(b) PCGrad

Figure 7: Impact of MTL on NAP accuracy in Production log (Transformer model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

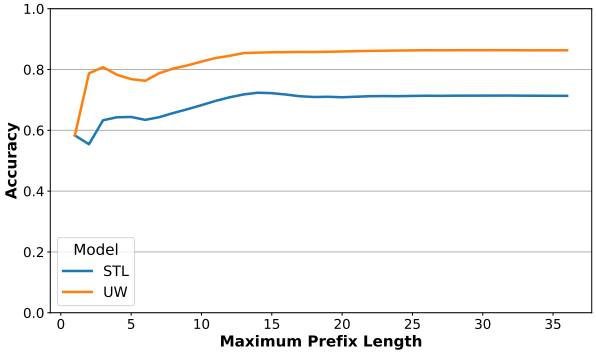


(a) UW

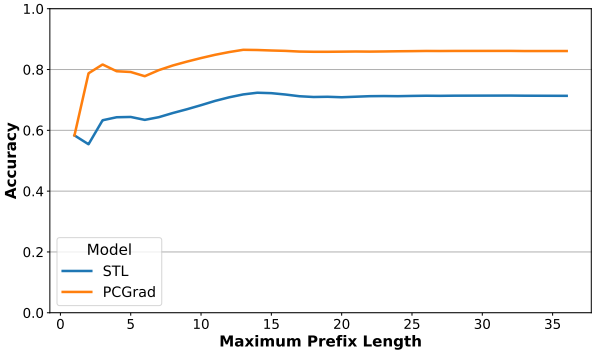


(b) PCGrad

Figure 8: Impact of MTL on NAP accuracy in P2P log (CNN model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

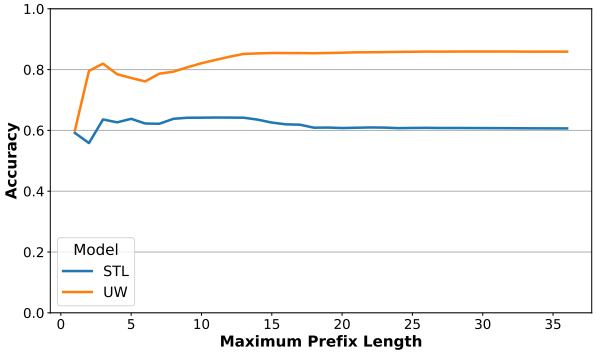


(a) UW

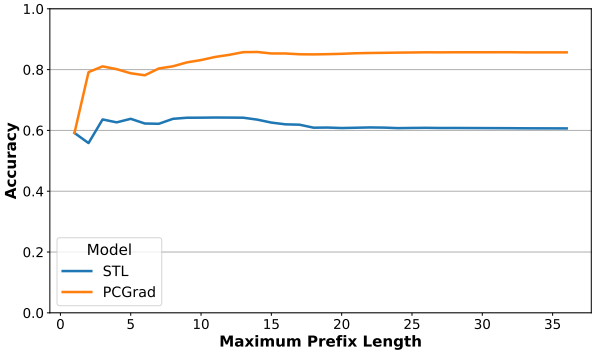


(b) PCGrad

Figure 9: Impact of MTL on NAP accuracy in P2P log (LSTM model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

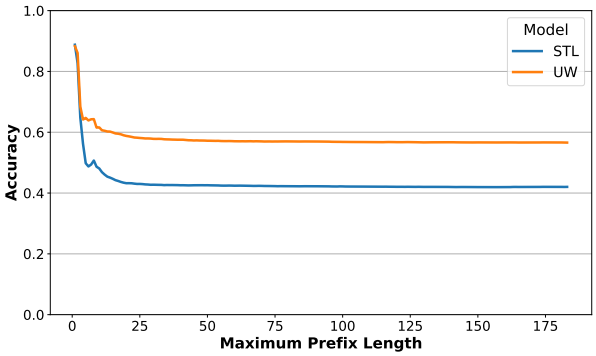


(a) UW

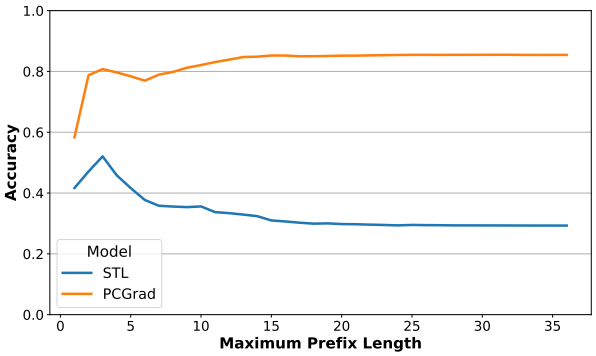


(b) PCGrad

Figure 10: Impact of MTL on NAP accuracy in P2P log (Transformer model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

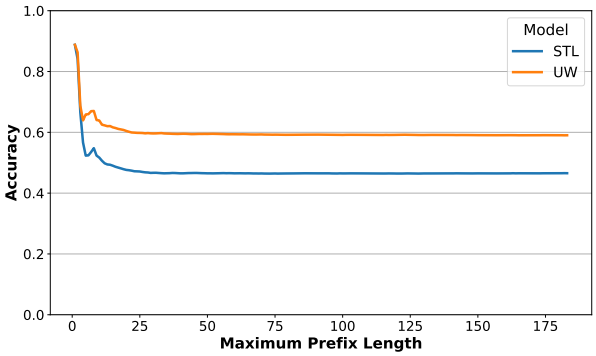


(a) UW

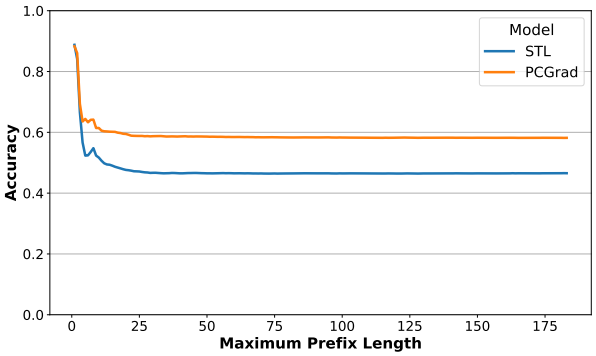


(b) PCGrad

Figure 11: Impact of MTL on NAP accuracy in Sepsis log (CNN model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

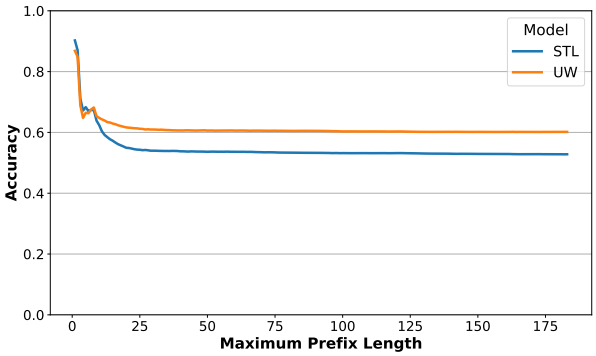


(a) UW

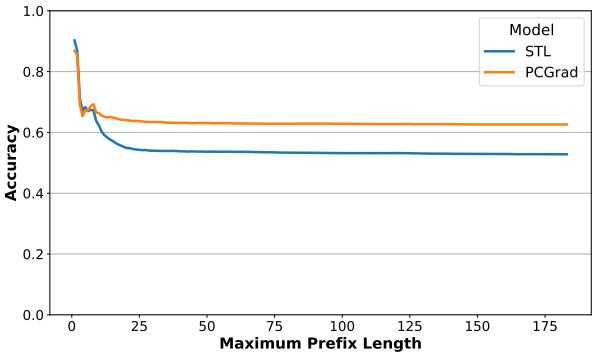


(b) PCGrad

Figure 12: Impact of MTL on NAP accuracy in Sepsis log (LSTM model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

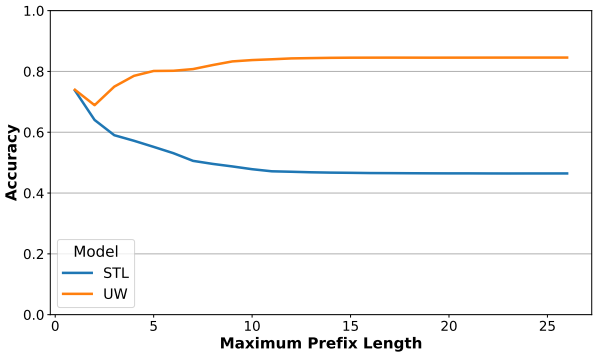


(a) UW

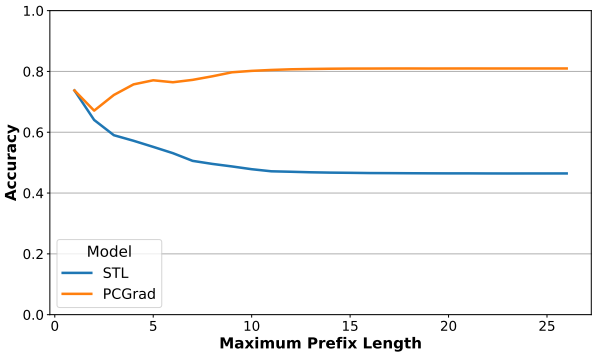


(b) PCGrad

Figure 13: Impact of MTL on NAP accuracy in Sepsis log (Transformer model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

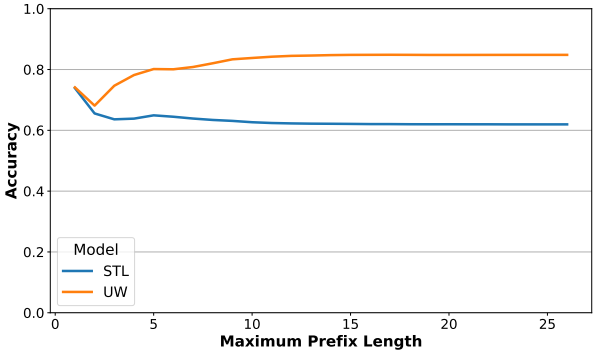


(a) UW

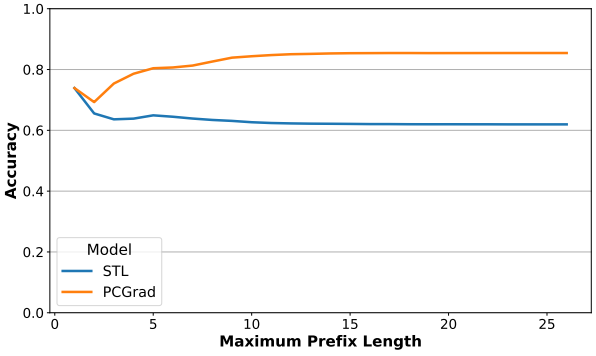


(b) PCGrad

Figure 14: Impact of MTL on NAP accuracy in BPIC20ID log (CNN model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

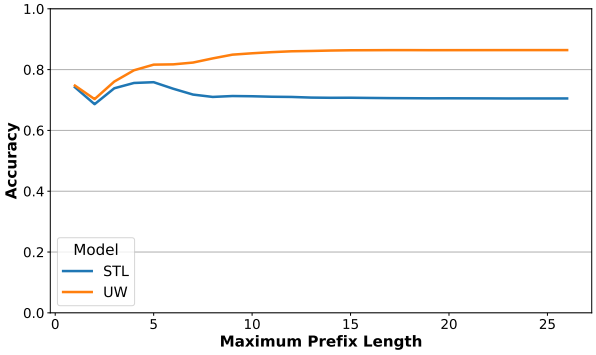


(a) UW

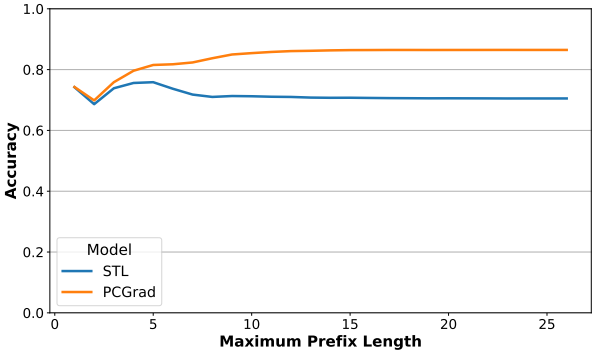


(b) PCGrad

Figure 15: Impact of MTL on NAP accuracy in BPIC20ID log (LSTM model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

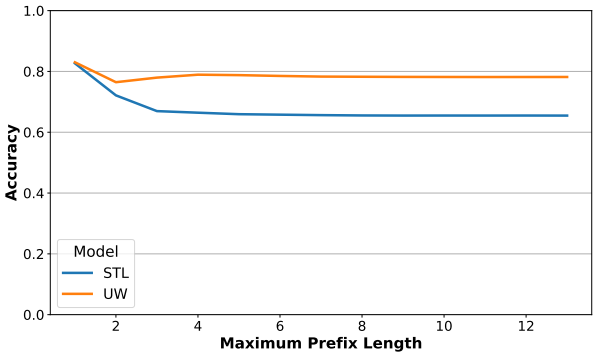


(a) UW

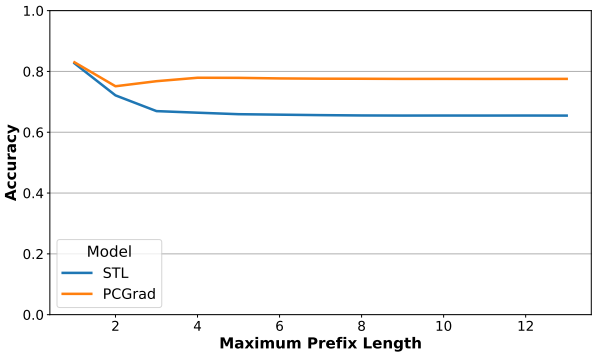


(b) PCGrad

Figure 16: Impact of MTL on NAP accuracy in BPIC20ID log (Transformer model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

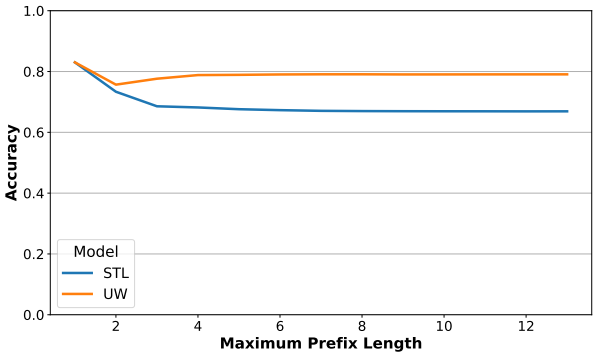


(a) UW

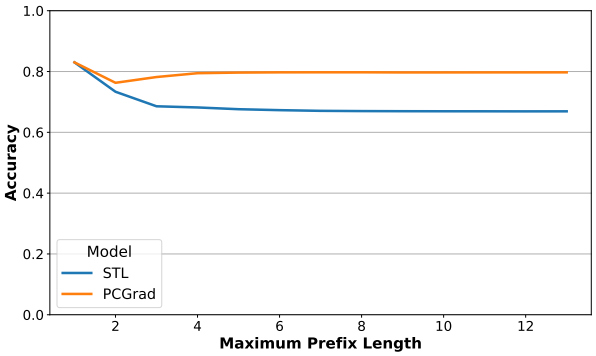


(b) PCGrad

Figure 17: Impact of MTL on NAP accuracy in HelpDesk log (CNN model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

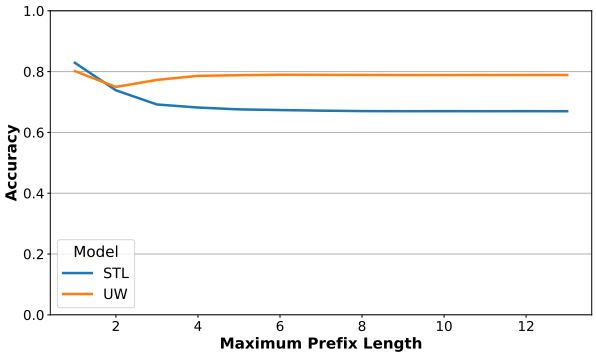


(a) UW

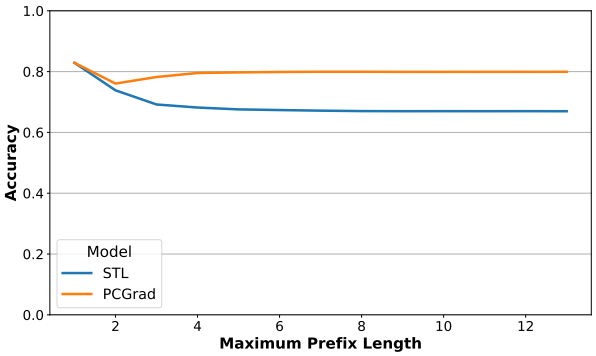


(b) PCGrad

Figure 18: Impact of MTL on NAP accuracy in HelpDesk log (LSTM model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.



(a) UW



(b) PCGrad

Figure 19: Impact of MTL on NAP accuracy in HelpDesk log (Transformer model with NAP+NTP+RTP task combination): prefix length analysis for (a) UW (b) PCGrad MTO methods.

References

Liu, B.; Liu, X.; Jin, X.; Stone, P.; and Liu, Q. 2021. Conflict-Averse Gradient Descent for Multi-task learning. In *NeurIPS*, 18878–18890.

Xin, D.; Ghorbani, B.; Garg, A.; Firat, O.; and Gilmer, J. 2022. Do Current Multi-Task Optimization Methods in Deep Learning Even Help? In *NeurIPS*.