Hierarchical Wasserstein Merging for Multi-Domain Multi-Task Learning: From Specialists to a Generalist

Ming Cheng[1], **Jiaying Gong**[2*], **Hoda Eldardiry**[1]
[1]Virginia Tech, [2]Amazon
{ming98,hdardiry}@vt.edu, gojiayin@amazon.com

Abstract

Multi-domain multi-task learning (MD-MTL) aims to build a single generalist model that performs well across heterogeneous domains and tasks. However, joint training often suffers from interference under distribution shifts. Existing model merging methods mostly operate on model parameters while overlooking the geometric structure of latent representation distributions across domains and tasks. To address these limitations, we propose Hierarchical Wasserstein Merging (HWM), a representation-level framework that models each domain-task specialist as a distribution of hidden representations on a shared support. HWM constructs task-level and global Wasserstein barycenters to capture within-task domain variation and cross-task structure, enabling either training-free specialist aggregation by Wasserstein-derived weights or training-based generalist learning through a hybrid Wasserstein alignment loss. Experiments on four NLP tasks across four domains per task show that HWM achieves superior effectiveness and generalization capability in MD-MTL settings.

1 Introduction

Multi-domain and multi-task learning (MD-MTL) aims to improve learning by sharing knowledge across heterogeneous domains and tasks. Learning one unified model that can handle multiple domains and tasks offers the potential for better generalization (Yang and Hospedales, 2015). However, jointly learning across domains and tasks remains challenging as differences in data distributions and linguistic patterns can lead to interference and negative transfer when domains or tasks are dissimilar.

The goal of MD-MTL is to integrate knowledge from a diverse set of specialists (domain and task-specific fine-tuned LLMs) into a single generalist model with broad multi-domain and multi-task capabilities. A straightforward solution is data merging, where the LLM is trained directly on the union of all available datasets. While simple, this approach requires access to large quantities of labeled data simultaneously, careful balancing of task objectives, and domain and task identifiers during inference. Recent work on model merging tends to integrate multiple specialist models into a unified model without extensive retraining. Existing merging techniques typically operate at the parameter level, task level, layer level, or neuron level, while overlooking how knowledge is encoded in the model's representation distributions. Directly interpolating or aligning parameters may fail to preserve the geometric structure of latent representations, particularly when specialists are trained on heterogeneous domains and tasks. This motivates a distributional view of MD-MTL model merging. In heterogeneous domain-task settings, the differences among specialists are not only reflected in individual model parameters, but also in how each specialist organizes inputs in its latent representation space. A single feature vector or a linear parameter average may obscure domain-specific structures, whereas a probability distribution over representations captures the geometry and spread of specialist behavior on a shared probing support.

There are three main challenges for MD-MTL merging. First is cross-domain geometry mismatch. Different domains can induce distinct latent structures, causing representation distributions to shift. The second challenge is cross-task hierarchy, where tasks may share partial structure (i.e., information extraction and question answering share factual grounding) while also requiring task-specific semantics. The third challenge is that a generalist is expected to integrate all specialists without requiring domain or task identifiers at inference time.

To address these challenges, we propose Hierarchical Wasserstein Merging (HWM), a

*This work was completed prior to joining Amazon, and does not relate to the author's position at Amazon

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 49457–49476
October 24-29, 2026 ©2026 Association for Computational Linguistics

representation-level framework for MD-MTL model merging through distributional alignment. HWM characterizes each domain-task specialist by the distribution of its hidden representations on a shared support. Using the entropically regularized Wasserstein distance, HWM constructs task-level barycenters that aggregate domains within each task and a global barycenter that captures shared structure across tasks. This hierarchical construction models both within-task domain diversity and cross-task relationships in a geometry-aware manner. By operating on representation distributions, HWM captures how each specialist organizes semantic information, enabling a more stable and interpretable aggregation mechanism. HWM supports two integration strategies: (1) For the non-training method, Wasserstein-derived hierarchical weights are used to aggregate specialists into a generalist without additional training. (2) For the training-based method, a generalist model is trained using supervised fine-tuning augmented with a Wasserstein alignment loss to align with pre-computed task-level and global barycenters. Both strategies operate without requiring domain or task identifiers during inference. The main contributions of this paper are summarized as:

- We propose Hierarchical Wasserstein Merging (HWM), a representation-level framework for multi-domain multi-task model merging based on distributional alignment.
- We develop non-training-based and training-based variants of HWM to support both specialists aggregation via Wasserstein-derived weights and generalist learning through a hybrid supervised and alignment loss.
- Extensive experiments across multiple tasks and domains demonstrate the superior effectiveness and generalization capability of HWM in multi-domain multi-task settings.

2 Related Work

2.1 Multi-Domain Multi-Task Learning

Multi-Domain Learning (MDL) aims to learn representations that generalize across domains through domain adaptation (Ding et al., 2025; Zhang et al., 2025b; Sun et al., 2025a; Reheman et al., 2025) and domain generalization (Yu et al., 2025; Liu et al., 2025a; Zhao et al., 2025a; Cheng et al., 2024). Multi-Task Learning (MTL) jointly optimizes related tasks to capture complementary knowledge (Chen et al., 2024). Existing MTL methods are typically architecture-based (Kong et al., 2025; Xu et al., 2025; Qi et al., 2025; Calvo et al., 2025; Zhu et al., 2024; Liu et al., 2024; Xu et al., 2024) and optimization-based methods (Zhang et al., 2025d; Li et al., 2025b; Hu et al., 2025; Gong et al., 2024). Recent studies combine MDL and MTL to exploit cross-domain and cross-task synergies (Zhang et al., 2024; Kwan et al., 2024; Xin et al., 2024). Inspired by recent advances in model merging, we introduce hierarchical Wasserstein merging, a geometry-aware framework for composing multi-domain multi-task specialists into a unified generalist model.

2.2 Model and Representation Merging

To integrate specialized knowledge from multiple sources, a data-centric strategy of data merging (Jiang et al., 2025; Corrado et al., 2025) trains a unified generalist model on the union of heterogeneous datasets. To enable the training-free integration of fine-tuned models, model merging (Sun et al., 2025b; Fu et al., 2025; Kirsch et al., 2025; Zhang et al., 2025c; Xie et al., 2025; Qiao et al., 2025) and LoRA merging (Ledoyen et al., 2025; Cheng et al., 2025a; Zhao et al., 2025b,c; Liu et al., 2025b; Zhang et al., 2025a; Zhang and Zhou, 2025) combine the capabilities of multiple fine-tuned models or adapters into a single model without additional task-specific training (Fu et al., 2025).

However, existing merging techniques primarily operate at parameter level (Li et al., 2025c; Qu et al., 2025; Yu et al., 2024; Jang et al., 2024; Yadav et al., 2023; Wortsman et al., 2022), task level (Qiu et al., 2025; Akiba et al., 2025; Yu et al., 2024; Davari and Belilovsky, 2024), layer level (Wang et al., 2025; Yang et al., 2024), or even neuron level (Fang et al., 2025). These approaches interpolate or align model weights while overlooking how knowledge is encoded in hierarchically organized representation distributions. Inspired by optimal transport (Imfeld et al., 2024; Singh and Jaggi, 2020), Hierarchical Wasserstein Merging merges models in representation distributions over specialist embeddings, enabling a generalist model to preserve hierarchical geometric consistency across domains and tasks.

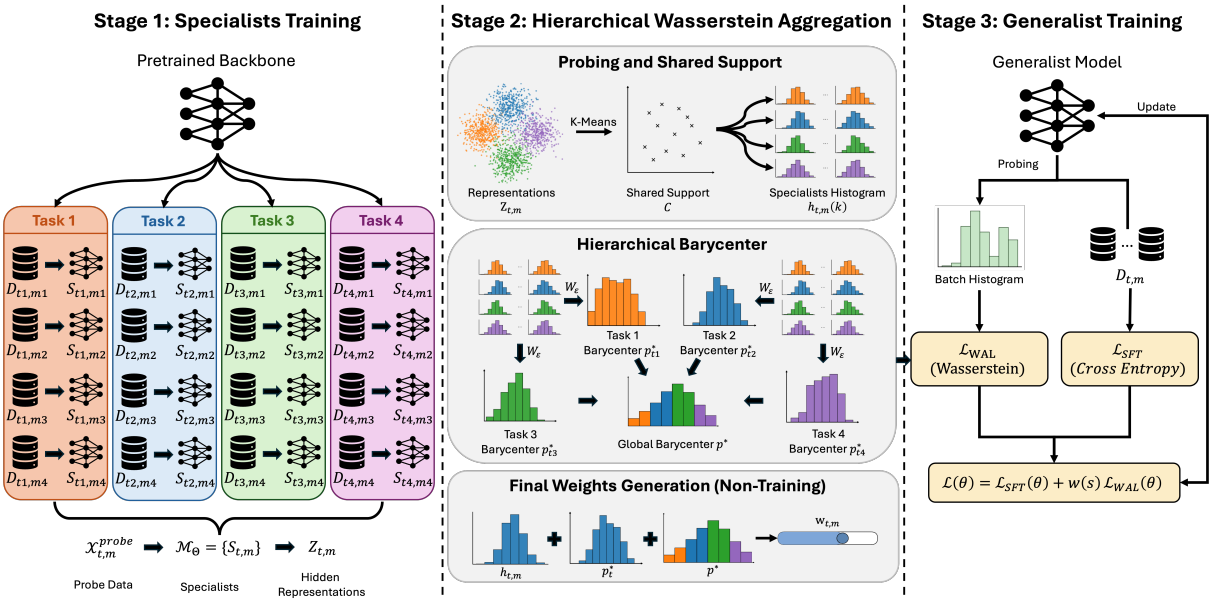

Figure 1: Overview of the proposed **Hierarchical Wasserstein Merging (HWM)** framework. **Stage 1:** Domain–task specialists ($S_{t,m}$) are fine-tuned from a shared pretrained backbone. **Stage 2:** Hidden representations are probed and aggregated on a shared support via entropic Wasserstein barycenters, resulting in task-level barycenters (p_t^*) and a global barycenter (p^*). **Stage 3:** A single generalist model is trained using a supervised loss ($\mathcal{L}_{\text{SFT}}$) augmented with a Wasserstein alignment loss ($\mathcal{L}_{\text{WAL}}$), which aligns the batch-level representation distribution (h_θ) to the precomputed task-level and global structures.

3 Methodology

3.1 Preliminaries

3.1.1 Wasserstein Distance and Barycenters

Wasserstein barycenters extend the classical notion of averaging from individual points to probability distributions or point clouds (Altschuler and Boix-Adsera, 2021). Given a set of probability distributions $\{\mu_i\}_{i=1}^n$, the Wasserstein barycenter is defined as the distribution ν that minimizes:

$$\nu = \arg\min_{\nu} \sum_{i=1}^{n} \lambda_i \, \mathcal{W}(\mu_i, \nu) \tag{1}$$

where $\{\lambda_i\}_{i=1}^n$ denotes non-negative weight satisfying $\sum_i \lambda_i = 1$, $\mathcal{W}(\cdot,\cdot)$ is the Wasserstein distance between distributions of hidden representations (Agueh and Carlier, 2011). The 2-Wasserstein distance can be defined as:

$$W_2^2(\mu,\nu) = \inf_{\gamma \in \Pi(\mu,\nu)} \mathbb{E}_{(x,y)\sim\gamma}\left[\|x-y\|_2^2\right] \tag{2}$$

where $\Pi(\mu,\nu)$ denotes the set of all joint probability measures with marginals μ and ν, $\|\cdot\|_2$ is the ℓ_2 norm, and $\mathrm{E}_{(x,y)\sim\gamma}[\cdot]$ denotes the expectation taken with respect to γ. In this paper, we compute a regularized approximation $\lim_{\varepsilon\to 0} W_\varepsilon(p,q) = W_2(p,q)$ for efficiency and stability.

3.1.2 Problem Formulation

Multi-domain multi-task learning (MD-MTL) aims to learn a single generalist model by integrating knowledge from multiple domains and task-specific specialists, where domains exhibit diverse data distributions and tasks differ in learning objectives. Formally, we consider a set of tasks $\mathcal{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_T\}$ and a set of domains $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_M\}$. Each task–domain pair $(\mathcal{T}_t, \mathcal{D}_m)$ is associated with a dataset $\mathcal{D}_{t,m} = \{(x_i^{t,m}, y_i^{t,m})\}_{i=1}^{N_{t,m}}$, where $x_i^{t,m}$ denotes the input and $y_i^{t,m}$ denotes the corresponding output. The goal of MD-MTL is to learn a single generalist model F_θ that simultaneously satisfies the requirements across all domains and tasks.

3.2 Hierarchical Wasserstein Merging

In this section, we describe our proposed HWM as shown in Figure 1 for models merging in multi-domain and multi-task scenarios. Our proposed method HWM contains three major stages: the specialists' fine-tuning stage (Sec. 3.2.1), the representation-level hierarchical Wasserstein aggregation stage (Sec. 3.2.2), and the single generalist training stage (Sec. 3.2.3)

3.2.1 Specialists Training

The proposed model consists of n domain-task specialists, where we primarily consider $n = 16$ in this paper across 4 tasks and 4 domains per task. The specialists are collectively denoted as:

$$\mathcal{M}_\Theta = \{S_{t,m} \mid t \in \mathcal{T},\ m \in \mathcal{D}_t\} \tag{3}$$

where $t \in \mathcal{T}$ denotes the task and $m \in \mathcal{D}_t$ denotes the domain associated with task t. Details of $\mathcal{T}$ and $\mathcal{D}_t$ are summarized in Table 1 in the Appendix A.2. Each specialist $S_{t,m}$ is independently fine-tuned on its corresponding domain-task dataset $\mathcal{D}_{t,m}$. We adopt Llama-3.1-8B-Instruct as the pretrained backbone for all $S_{t,m}$ due to its strong instruction-following and generalization ability (Dubey et al., 2024).

3.2.2 Representation-Level Wasserstein Aggregation

Probing Distribution and Shared Support. To enable representation-level comparison across specialists, we first probe their internal activations. For each specialist $S_{t,m}$, we collect a probe set $\mathcal{X}^{\text{probe}}_{t,m}$ and extract hidden representations from the final transformer layer, which captures high-level semantic information. Given an input sequence $x \in \mathcal{X}^{\text{probe}}_{t,m}$, let $\phi_\ell(x) = \{h_1, \ldots, h_{L_x}\}$ denote the sequence of hidden states at layer ℓ. The probed representation set for each specialist is:

$$Z_{t,m} = \left\{ \frac{1}{L_x} \sum_{i=1}^{L_x} h_i \;\middle|\; x \in \mathcal{X}^{\text{probe}}_{t,m} \right\} \tag{4}$$

where L_x is the number of tokens in x. We further do dimensionality reduction via principal component analysis to facilitate scalable clustering and optimal transport computation. Next, we construct a shared discrete support over the representation space by applying k-means clustering to get cluster centers $\mathcal{C} = \{c_1, \ldots, c_K\}$. Each specialist is then represented as a probability distribution over the shared support. We compute a soft histogram $h_{t,m}(k)$ as the empirical distributional representation of specialist $S_{t,m}$ on the shared support $\mathcal{C}$:

$$h_{t,m}(k) = \frac{1}{|Z_{t,m}|} \sum_{z \in Z_{t,m}} \frac{\exp\big(-\|z - c_k\|_2^2/(2\tau^2)\big)}{\sum_{j=1}^{K} \exp(-\|z - c_j\|_2^2/(2\tau^2))} \tag{5}$$

where $\tau > 0$ controls the smoothness, and $k \in \{1, \cdots K\}$ indexes the cluster centroids.

Hierarchical Wasserstein Aggregation. To aggregate domain-task distributions, we compute entropic Wasserstein barycenters on the shared support $\mathcal{C}$. Given histogram representations $\{h_{t,m}\}$ and entropic kernel $K \in \mathbb{R}^{K \times K}$, where $K_{ij} = \exp(-\|c_i - c_j\|_2^2/\varepsilon)$, we first compute a task-level barycenter p_t^* by aggregating all domain distributions associated with the same task t:

$$p_t^* = \arg\min_{p \in \mathcal{P}(\mathcal{C})} \sum_{m \in \mathcal{M}_t} W_\varepsilon(p, h_{t,m}) \tag{6}$$

where $W_\varepsilon(\cdot,\cdot)$ denotes the entropic Wasserstein distance computed with kernel K. We aggregate task-level barycenters p_t^* to obtain a global barycenter p^* that integrates information across all tasks:

$$p^* = \arg\min_{p \in \mathcal{P}(\mathcal{C})} \sum_{t \in \mathcal{T}} W_\varepsilon(p, p_t^*) \tag{7}$$

Both barycenters are computed using Sinkhorn fixed-point iterations on the shared support, leading to stable and scalable solutions. For a non-training-based HWM method, the final weight $w_{t,m}$ assigned to the specialist model $S_{t,m}$ is obtained by combining task-level and within-task relevance through a hierarchical weighting method:

$$w_{t,m} = \frac{\exp\big(-\beta_{\text{task}}\, W_\varepsilon(p_t^*, p^*)\big)}{\sum_{t' \in \mathcal{T}} \exp\big(-\beta_{\text{task}}\, W_\varepsilon(p_{t'}^*, p^*)\big)} \times \frac{\exp\big(-\beta_{\text{in}}\, W_\varepsilon(h_{t,m}, p_t^*)\big)}{\sum_{m' \in \mathcal{M}_t} \exp\big(-\beta_{\text{in}}\, W_\varepsilon(h_{t,m'}, p_t^*)\big)} \tag{8}$$

where $\sum_{t \in \mathcal{T}} \sum_{m \in \mathcal{M}_t} w_{t,m} = 1$, β_{task} and β_{in} control the sharpness of task-level and within-task selection, and $W_\varepsilon(\cdot,\cdot)$ denotes the entropically regularized Wasserstein distance computed on the shared support. This final weight prioritizes specialists whose representation distributions are geometrically closer to the corresponding task-level barycenter and the global barycenter.

3.2.3 Generalist Training

We train a single generalist model using standard supervised fine-tuning augmented with a representation-level Wasserstein alignment loss. The generalist model is initialized from the same pretrained backbone as specialists $\mathcal{M}_\Theta$ and optimized over labeled multi-domain and multi-task data. The overall training objective is defined as:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{SFT}}(\theta) + w(s)\, \mathcal{L}_{\text{WAL}}(\theta) \tag{9}$$

where θ denotes the parameters of the generalist model, $\mathcal{L}_{\text{SFT}}(\theta)$ is the standard supervised loss,

Table 1: Overview of the 30 publicly available datasets used across four NLP tasks in multiple domains.

Task	Domain	Datasets Used
Paraphrasing	*Biomedical*	eLife (Goldsack et al., 2022), PLOS (Goldsack et al., 2022), Cells (Guo et al., 2024)
	News	SciTechNews (Cardenas et al., 2023), MSRP (Dolan and Brockett, 2005)
	Science	VTechAGP (Cheng et al., 2025b), ParaSci (Dong et al., 2021)
	Wikipedia	PAWS (Zhang et al., 2019)
Question Answering	*Web*	ComplexWebQuestions (Talmor and Berant, 2018), WebQuestions (Berant et al., 2013), TriviaQA (Joshi et al., 2017)
	Biomedical	COVID-QA (Möller et al., 2020), HealthQA (Zhu et al., 2019)
	Wikipedia	WikiQA (Yang et al., 2015), Natural Questions (Kwiatkowski et al., 2019)
	News	NewsQA (Trischler et al., 2017)
Summarization	*News*	Newsroom (Grusky et al., 2018), Multi-News (Fabbri et al., 2019), WCEP-MDS (Gholipour Ghalandari et al., 2020)
	Literature	BookSum (Kryscinski et al., 2022), SQuALITY (Wang et al., 2022)
	Government	BillSum (Kornilova and Eidelman, 2019), BigPatent (Sharma et al., 2019), GovReport (Huang et al., 2021)
	Science	ArXiv & PubMed (Cohan et al., 2018), SciSummNet (Yasunaga et al., 2019), BigSurvey (LIU et al., 2022), Multi-XScience (Lu et al., 2020), SumPubMed (Gupta et al., 2021)
Information Extraction	*News*	IEPile (DuIE2.0, NYT, NYT11-HRL, CoNLL2004, KBP37, GIDS, CLUE, MSRA, ACE2005, CoNLL2003, DuEE1.0)
	Biomedical	IEPile (CMeIE, ADE Corpus, PHEE, AnatEM, BC2GM, BC4CHEMD, BC5CDR, NCBI-Disease, GENIA)
	Science	IEPile (SemEval RE, SciERC, FabNER, CASIE)
	Wikipedia	IEPile (MultiNERD) (Gui et al., 2024)

and $\mathcal{L}_{\text{WAL}}(\theta)$ is the Wasserstein loss for aligning the generalist with both task-specific and global distributional structures. $w(s)$ is a scheduling function of the optimization step s that gradually increases the influence of the Wasserstein loss after an initial warm-up period, which stabilizes optimization by allowing the model to first learn task-relevant supervision before enforcing distributional alignment. Specifically, $\mathcal{L}_{\text{SFT}}(\theta)$ is the cross-entropy loss computed on labeled training data, and $\mathcal{L}_{\text{WAL}}(\theta)$ is a weighted combination of task-level and global Wasserstein alignment losses:

$$\mathcal{L}_{\text{WAL}}(\theta) = \lambda_{\text{task}}\, \mathcal{L}_{\text{WAL}}^{\text{task}}(\theta) + \lambda_{\text{global}}\, \mathcal{L}_{\text{WAL}}^{\text{global}}(\theta) \tag{10}$$

where λ_{task}, and λ_{global} are the weighting parameters. Specifically, $\mathcal{L}_{\text{WAL}}^{\text{task}}(\theta) = W_{\varepsilon}(h_{\theta}, p_t^*)$, and $\mathcal{L}_{\text{WAL}}^{\text{global}}(\theta) = W_{\varepsilon}(h_{\theta}, p^*)$, where h_{θ} is the batch-level representation histogram.

4 Experiments

4.1 Experiment Settings

4.1.1 Datasets

We evaluate our proposed model on four representative NLP tasks: Paraphrasing, Summarization, Question Answering, and Information Extraction. Each task covers four distinct domains, as summarized in Table 1, which lists the corresponding datasets used for each domain. In total, our experiments involve 30 publicly available datasets for training and evaluation. To ensure fair comparison and reproducibility, we follow the same training, validation, and testing splits provided by each dataset. Additionally, we construct a probing set for each task and domain by randomly sampling 2k instances from the original training set, ensuring that the probing data do not overlap with any examples used for training, validation, or testing. We further sample 5k data from the training set for each task and domain for the generalist training.

4.1.2 Baselines

We compare HWM against several representative baselines. These include the pretrained base model (LLaMA3.1-8B-Instruct (Dubey et al., 2024)), specialists (models fine-tuned based only on individual domain-task datasets), and a generalist model fine-tuned jointly on the union of all domain-task data. For model merging approaches, we consider a range of general model merging methods, including Model Soup (Wortsman et al., 2022), Task Arithmetic (Ilharco et al., 2023), TIES (Yadav et al., 2023), DARE (Yu et al., 2024), DELLA (Deep et al., 2024), Model Breadcrumbs (Davari and Belilovsky, 2024), and Model Stock (Jang et al., 2024). For our approach, we evaluate two variants of Hierarchical Wasserstein Merging: (i) HWM, a non-training-based version that aggregates specialists using Wasserstein-derived hierarchical weights, and (ii) HWM*, a training-based single generalist learned through a hybrid Wasserstein loss.

4.1.3 Evaluation Metrics

We adopt task-specific evaluation metrics to assess model performance across different tasks. Overall, our evaluation metrics consist of two categories: automative model assessment and LLM-as-a-judge, where GPT4.1 is used to score the candidate outputs given the corresponding input context and task instructions (Li et al., 2025a). LLM-as-a-judge

Table 2: Performance (%) comparison of merging methods for text summarization across four domains (Government, Literature, News, and Science). We report BERTScore F1 (%), SARI, and human-aligned preference scores using LLM-as-a-judge. The overall best result is in bold, and the best merging result is underlined.

Types	Methods	BERTScore F1 (%)				SARI				LLM-as-a-judge (%)			
		Gov.	Lit.	News	Sci.	Gov.	Lit.	News	Sci.	Gov.	Lit.	News	Sci.
Pre-trained	Pretrained	74.92	73.18	71.40	72.63	35.21	36.08	36.44	34.97	52.34	42.10	62.93	72.53
Fine-tuned	Generalist	81.68	76.22	74.33	80.49	44.88	39.22	37.40	44.76	51.76	39.87	57.56	63.54
	Specialist	**82.12**	77.25	76.58	**81.27**	**45.56**	39.70	41.50	**45.27**	53.26	44.14	65.69	69.84
Merged	Model Soup	76.81	76.49	75.39	76.93	40.32	37.47	35.33	38.52	49.96	44.38	59.20	63.88
	Task Arithmetic	78.43	76.72	76.30	75.87	38.53	39.16	38.78	37.78	64.22	50.46	60.51	72.27
	TIES	78.03	74.41	74.34	75.20	37.48	38.07	40.44	37.74	58.98	35.49	48.57	65.26
	DARE	77.66	74.75	75.75	75.40	38.13	38.18	38.59	37.79	59.14	35.44	54.46	66.90
	DELLA	78.02	74.41	74.74	75.21	38.26	38.07	40.29	38.63	58.10	35.66	50.70	67.02
	Breadcrumbs	77.03	76.31	75.25	75.67	37.23	38.94	40.23	37.04	64.37	48.32	59.39	72.09
	Model Stock	65.93	66.02	64.68	64.53	33.78	37.88	38.36	35.65	59.24	35.07	46.60	56.38
	HWM (ours)	79.94	76.53	77.30	77.75	41.15	39.54	42.09	39.61	**67.28**	**53.32**	60.79	78.40
	HWM* (ours)	77.28	**77.41**	**77.37**	79.48	39.48	**39.96**	**42.24**	40.21	63.00	53.23	**66.13**	**80.90**

evaluation is applied to all four tasks [1], and the detailed prompts are provided in Appendix A.3. For text summarization and lay paraphrasing, we report BERTScore (F1) (Zhang* et al., 2020), which measures semantic similarity between generated and reference texts, and SARI (Xu et al., 2016), which evaluates content selection and rewriting behavior. For the generative question answering task, we use BERTScore (F1) (Zhang* et al., 2020) and METEOR (Banerjee and Lavie, 2005), as METEOR better captures semantic matches in generative QA and is more suitable for abstractive responses. For information extraction, we report the prediction validity rate, defined as the percentage of model outputs that can be parsed into a valid dictionary, to measure format reliability. We also report Micro-F1, computed as the micro-averaged extraction F1 score over canonicalized outputs using soft matching, to assess extraction accuracy.

4.1.4 Parameter Settings

We implement Hierarchical Wasserstein Merging (HWM) in Pytorch and finetune both specialists and generalist models using LLaMA-3.1-8B-Instruct (Dubey et al., 2024) by LLaMA-Factory [2] as the shared pretrained backbone. For specialists training, we use the learning rate of 1e-5 and train each specialist for one epoch on its corresponding domain-task dataset $\mathcal{D}_{t,m}$. For Wasserstein aggregation, we sample up to 2k probing examples per dataset.

The number of cluster centroids K is selected by grid search from $\{128, 256, 512\}$. Histogram smoothness τ is 1.2 in equation 5. The entropic regularization ε is 0.07 for building the kernel K_{ij}. Barycenters run for up to 200 iterations with a tolerance of 10^{-9}. For the non-training hierarchical weighting in Equ. 8, β_{task} and β_{in} are 0.1 and 0.5, respectively. For generalist training, both λ_{task} and λ_{global} are 0.3 in Equ. 10. The Sinkhorn solver for the alignment loss runs for 10 iterations, and the WAL loss is applied periodically every two optimization steps. All merging baselines are performed using mergekit [3].

4.2 Results and Analysis

4.2.1 Main Results

Text Summarization. Table 2 reports summarization results across four domains using BERTScore F1, SARI, and LLM-as-a-judge. HWM achieves the strongest overall performance among merging methods, obtaining the best or second-best results on BERTScore and SARI in the Literature, News, and Science domains. It also substantially improves LLM-as-a-judge scores, especially in Science, suggesting that representation-level Wasserstein aggregation better preserves summary quality aligned with human preferences. While fully fine-tuned generalists or specialists can be competitive in data-rich domains such as Government and Science, HWM remains robust across domains and shows clear advantages under greater domain heterogeneity.

Lay Paraphrasing. Table 3 reports lay paraphrasing results across four domains using BERTScore F1, SARI, and LLM-as-a-judge. Among merging-based approaches, HWM consistently achieves the strongest overall performance,

[1] Due to the cost of GPT4.1, we randomly sample 1k testing data per domain of each task for LLM-as-a-judge assessment.

[2] https://github.com/hiyouga/LlamaFactory

[3] https://github.com/arcee-ai/mergekit

Table 3: Performance (%) comparison of merging methods for lay paraphrasing across four domains (Biology, News, Science, and Wikipedia). We report BERTScore F1 (%), SARI, and human-aligned preference scores using LLM-as-a-judge. The overall best result is in bold, and the best merging result is underlined.

Types	Methods	BERTScore F1 (%)				SARI				LLM-as-a-judge (%)			
		Bio.	News	Sci.	Wiki	Bio.	News	Sci.	Wiki	Bio.	News	Sci.	Wiki
Pre-trained	Pretrained	70.16	61.20	62.81	69.25	38.24	23.80	25.65	28.23	54.38	10.26	9.19	24.71
Fine-tuned	Generalist	77.18	69.85	74.58	78.15	38.75	40.26	34.48	42.44	55.28	14.59	10.09	20.03
	Specialist	81.41	74.08	**81.19**	**90.60**	41.46	**44.99**	**39.33**	**51.79**	61.61	21.21	33.51	**65.06**
Merged	Model Soup	81.08	76.64	78.88	84.34	39.36	36.43	34.80	34.84	53.77	25.36	15.17	40.86
	Task Arithmetic	80.40	74.59	75.45	75.20	38.70	33.86	34.89	33.86	64.56	23.52	26.61	21.07
	TIES	80.46	78.21	78.88	84.75	38.45	33.81	33.42	21.99	69.91	39.52	22.48	40.28
	DARE	81.10	78.21	77.56	78.19	38.72	33.81	34.52	34.81	68.46	42.86	39.71	33.02
	DELLA	80.47	78.59	79.09	76.43	38.50	33.79	33.41	35.03	66.19	41.33	31.41	22.21
	Breadcrumbs	80.58	74.59	80.26	80.93	39.92	33.86	30.86	21.57	64.40	23.42	25.62	21.10
	Model Stock	65.98	59.52	65.77	62.56	34.26	26.90	28.52	27.52	52.70	22.81	13.44	40.70
	HWM (ours)	81.74	79.27	80.65	83.33	40.19	38.92	34.52	36.87	**70.14**	43.49	**43.71**	44.77
	HWM* (ours)	**83.48**	**79.84**	79.12	77.70	**42.48**	40.97	32.39	33.93	64.37	**45.06**	36.49	46.48

Table 4: Performance (%) comparison of merging methods for generative question answering across four domains (Biology, News, Web, and Wikipedia). We report BERTScore F1 (%), METEOR (%), and human-aligned preference scores using LLM-as-a-judge. The overall best result is in bold, and the best merging result is underlined.

Types	Methods	BERTScore F1 (%)				METEOR (%)				LLM-as-a-judge (%)			
		Bio.	News	Web	Wiki	Bio.	News	Web	Wiki	Bio.	News	Web	Wiki
Pre-trained	Pretrained	67.66	72.92	67.51	67.36	12.49	12.54	20.32	15.85	25.67	25.02	58.70	70.87
Fine-tuned	Generalist	72.54	77.53	85.40	69.74	18.69	31.26	18.31	15.38	48.15	41.15	59.57	61.50
	Specialist	77.43	77.77	**85.66**	72.04	22.17	**32.01**	**36.25**	**28.16**	64.83	47.17	68.11	74.85
Merged	Model Soup	74.34	77.50	72.46	75.85	11.57	30.81	23.40	23.03	46.26	43.47	58.68	66.37
	Task Arithmetic	75.12	75.56	73.50	76.80	16.49	23.49	27.71	27.65	63.02	**50.45**	**68.22**	79.57
	TIES	76.49	76.09	74.32	77.46	17.83	25.52	27.72	26.27	51.10	41.79	55.13	69.74
	DARE	74.39	76.16	73.36	77.45	12.28	25.77	27.25	27.15	54.46	44.77	63.29	79.40
	DELLA	75.78	77.09	75.27	77.49	18.46	25.60	27.72	26.31	59.93	42.86	55.36	69.30
	Breadcrumbs	75.88	76.67	73.50	76.09	**22.24**	26.86	27.71	26.28	**68.37**	49.19	67.90	79.86
	Model Stock	74.43	77.42	71.24	77.65	11.47	31.07	22.89	27.59	56.48	45.12	61.77	**81.64**
	HWM (ours)	**77.82**	77.42	77.80	**78.30**	21.23	28.21	28.45	27.11	67.17	45.40	62.62	79.69
	HWM* (ours)	77.64	**80.39**	79.54	78.00	19.93	31.50	30.91	27.65	65.13	49.51	67.19	81.18

getting the best or second-best results across most domains and metrics. In particular, HWM yields the highest BERTScore and SARI scores in the Biomedical and News domains, indicating improved semantic preservation and paraphrasing quality under domain shift. HWM also outperforms other merging baselines on LLM-as-a-judge scores in News and Wikipedia, suggesting better alignment with human preferences. While fully fine-tuned specialists are very competitive, especially in high-resource domains (i.e., Wikipedia), HWM substantially narrows the performance gap without requiring domain-specific inference, demonstrating its effectiveness as a generalist merging framework for lay paraphrasing.

Generative Question Answering. Table 4 summarizes generative question answering performance across four domains using BERTScore F1, METEOR, and LLM-as-a-judge. Among merging-based approaches, HWM achieves the strongest and most consistent performance across domains, obtaining the best merging results on BERTScore in all four domains and competitive METEOR scores, particularly in Biology, News, and Wikipedia. In addition, we observe that in Web and Wikipedia domains, the pretrained model already exhibits strong performance, likely due to broad exposure to similar content during pretraining. As a result, additional fine-tuning or merging yields more modest gains in these high-resource domains, while HWM shows clearer improvements in domains with greater distributional divergence.

Information Extraction. Table 5 presents information extraction results across four domains using prediction validity rate, Micro-F1, and LLM-as-a-judge. From Table 5, we observe that the jointly trained generalist performs poorly on information extraction, likely due to interference from heterogeneous tasks that disrupt the strict output structure required by IE. This behavior is similar to the pretrained model, which often fails to produce valid structured outputs, highlighting the challenge of

Table 5: Performance (%) comparison for information extraction across four domains (Biology, News, Science, and Wikipedia). We report output validity rate (%), Micro-F1 (%) for extraction accuracy, and human-aligned preference scores using LLM-as-a-judge. The overall best result is in bold, and the best merging result is underlined.

Types	Methods	Prediction Validity Rate (%)				Micro-F1 (%)				LLM-as-a-judge (%)			
		Bio.	News	Sci.	Wiki	Bio.	News	Sci.	Wiki	Bio.	News	Sci.	Wiki
Pre-trained	Pretrained	16.22	25.70	26.37	28.98	5.84	20.80	5.36	22.85	11.60	24.38	14.07	60.90
Fine-tuned	Generalist	3.55	8.20	8.16	15.36	0.44	8.69	7.10	19.02	24.70	17.66	15.32	59.15
	Specialist	47.19	92.67	55.02	84.00	**27.17**	31.01	11.40	33.33	**31.42**	44.05	17.66	62.40
Merged	Model Soup	25.08	16.05	22.29	16.45	9.36	16.07	5.70	18.17	15.36	15.18	15.40	53.16
	Task Arithmetic	14.85	97.65	17.99	**97.71**	5.37	32.84	6.87	37.65	19.81	35.92	13.64	67.50
	TIES	81.25	93.49	75.57	86.71	13.32	29.32	12.37	36.10	13.85	29.02	12.97	62.75
	DARE	30.03	90.30	77.83	87.25	8.67	27.32	12.52	31.04	20.50	28.37	14.42	64.15
	DELLA	81.58	90.35	26.41	87.36	11.49	27.35	9.21	31.11	17.67	28.05	12.24	64.54
	Breadcrumbs	**91.53**	98.05	**94.32**	97.71	13.43	32.60	**15.42**	37.68	14.58	36.81	13.42	**71.47**
	Model Stock	20.18	21.25	17.74	22.77	7.58	14.47	3.61	17.56	14.92	22.59	12.76	60.27
	HWM (ours)	85.71	98.25	94.10	94.12	12.34	36.60	15.27	37.14	22.90	44.87	20.48	69.27
	HWM* (ours)	85.66	**98.90**	92.20	95.56	11.81	**41.32**	15.38	**39.66**	23.65	**47.22**	**22.98**	70.39

Table 6: Ablation study of HWM components. Results are averaged across domains for each task and reported in percentages. BS denotes BERTScore F1, MET. denotes METEOR, Valid. denotes prediction validity rate, MiF1 denotes Micro-F1, and LLM denotes LLM-as-a-judge.

Types	Methods	Summarization			Lay			QA			IE		
		BS	SARI	LLM	BS	SARI	LLM	BS	MET.	LLM	Valid.	MiF1	LLM
Architecture	Global	74.98	37.03	57.31	80.89	34.46	46.76	76.31	24.11	60.62	76.38	15.88	22.35
	HWM	**77.88**	**40.60**	**64.95**	**81.25**	**37.63**	**50.53**	**77.84**	**26.25**	**63.72**	**93.05**	**25.34**	**39.38**
Centroids	HWM-128	75.28	38.50	51.27	80.35	34.00	47.88	76.01	38.64	60.47	89.25	19.95	34.12
	HWM-256	76.58	39.68	57.08	80.89	33.92	48.46	76.38	**38.66**	59.88	92.09	20.01	35.87
	HWM-512	77.88	**40.60**	64.95	**81.25**	**37.63**	**50.53**	77.84	26.25	63.72	93.05	25.34	39.38
	HWM*-128	73.14	36.57	60.68	78.71	35.35	43.35	76.73	37.39	60.12	90.48	23.54	39.63
	HWM*-256	75.56	38.29	61.87	79.04	36.48	45.50	77.73	37.86	64.12	92.50	25.81	37.93
	HWM*-512	**77.89**	40.47	**65.82**	80.04	37.44	48.10	**78.89**	27.50	**65.75**	**93.08**	**27.04**	**41.06**
Probing	HWM-small	75.13	38.40	60.48	79.60	34.30	48.68	75.39	24.12	56.61	87.27	20.97	34.94
	HWM	77.88	**40.60**	64.95	**81.25**	**37.63**	**50.53**	77.84	26.25	63.72	93.05	25.34	39.38
	HWM*-small	75.88	38.46	64.77	79.75	35.20	45.92	78.47	23.89	63.79	90.05	25.86	39.60
	HWM*	**77.89**	40.47	**65.82**	80.04	37.44	48.10	**78.89**	**27.50**	**65.75**	**93.08**	**27.04**	**41.06**

preserving task-specific output formats under joint training. Compared to other merging baselines that often exhibit unstable behavior across domains, HWM provides a more balanced performance profile, yielding competitive validity, higher Micro-F1, and strong human-aligned preference scores.

4.2.2 Ablation Study

To evaluate key design choices in HWM, we conduct ablation studies on the hierarchical architecture, the number of cluster centroids, and the probing-set size. Table 6 reports domain-averaged scores for each task. We observe that each component contributes positively. The hierarchical architecture consistently outperforms the global variant, with particularly large gains on the IE task. Increasing the number of centroids generally improves performance, with $K = 512$ achieving the best or near-best results in most cases. The full probing setup also outperforms the smaller probing set, especially on QA and IE tasks. These results support our choices of hierarchical aggregation, $K = 512$, and using the full probing set.

Effect of hierarchical architecture. Table 7 in the Appendix shows that the hierarchical architecture improves over the global architecture in nearly all domains and metrics. For summarization, HWM improves BERTScore and LLM-as-a-judge scores across all four domains, with especially large gains on the Science domain. Similar trends are observed for Lay and QA, where HWM improves most domain-level scores despite a few small decreases on individual metrics. The improvement is most consistent on IE where HWM outperforms the global architecture in every domain for prediction validity, Micro-F1, and LLM-as-a-judge scores. These results suggest that hierarchical aggregation better preserves domain-specific information before global fusion, leading to more robust performance across tasks and domains.

Effect of the number of centroids. Table 8 in the Appendix reports the effect of varying the number of cluster centroids K. Overall, increasing K

generally improves performance for both HWM and HWM*, with $K = 512$ achieving the best or near-best results in most domains. The gains are particularly clear for summarization and IE, where larger centroid sets improve most metrics across domains. Although a few individual domain-metric pairs show non-monotonic trends, the overall pattern indicates that finer centroid granularity helps capture more task- and domain-specific structure. This supports our choice of using $K = 512$ as the default setting.

Effect of probing size. Table 9 in Appendix compares the full probing setup with a smaller probing set. The full probing setup consistently improves HWM across most domains, especially for Summarization, Lay, and IE. For IE, using the full probing set improves prediction validity and Micro-F1 in every domain for both HWM and HWM*. QA also benefits from more probing examples, particularly in LLM-as-a-judge scores. While a small number of domain-level metrics decrease, the overall trend shows that additional probing examples provide more reliable task- and domain-level signals, leading to stronger model merging performance.

4.2.3 Additional Analysis

Non-training-based v.s. Training-based HWM Across all four tasks, both training (HWM*) and non-training (HWM) versions of our proposed model demonstrate strong performance as shown from Table 2 to Table 5. HWM consistently improves over existing training-free merging baselines, achieving competitive or best merging performance on semantic metrics such as BERTScore and LLM-as-a-judge scores, while requiring no additional optimization. HWM* further enhances performance by aligning the generalist's representations with task-level and global Wasserstein barycenters, leading to additional gains on metrics such as SARI for summarization and Micro-F1 for information extraction. In particular, HWM* shows clear advantages in domains where preserving task-specific structure is critical, while HWM remains robust and effective in scenarios where retraining is infeasible. These results highlight the flexibility of HWM, enabling both efficient training-free aggregation and performance-oriented generalist learning within a unified framework.

Task Weight Distributions. Figure 2 visualizes the task-level weight distributions produced by HWM for IE and QA inputs across multiple domains. For IE inputs, HWM consistently assigns higher weights to IE specialists, while still allocating weights to related tasks such as QA, reflecting shared factual and structural information. Similarly, for QA inputs, QA specialists receive dominant weights. These results demonstrate that HWM does not rely on hard task assignments. Instead, it performs fine-grained, sample-dependent weighting that captures both task alignment and cross-task relatedness, supporting its effectiveness as a flexible multi-domain multi-task merging framework.

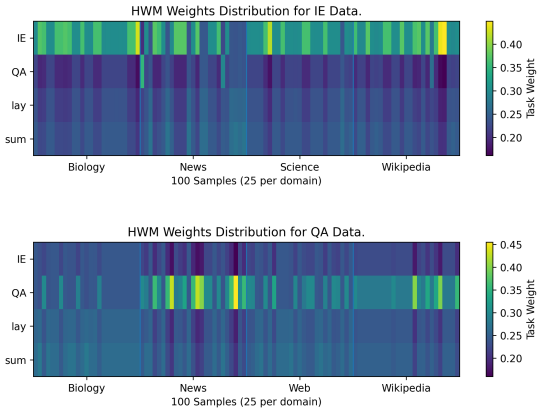

Figure 2: Visualization of HWM weights for IE and QA tasks. Each column is one data sample, and each row shows the average weight assigned to specialists from different tasks. Color intensity indicates the magnitude of the assigned task weight.

5 Conclusion

In this paper, we propose Hierarchical Wasserstein Merging (HWM), a unified framework for multi-domain multi-task learning. Without requiring domain and task-specific information during inference, HWM supports both non-training-based specialist aggregation through Wasserstein-derived weights and training-based generalist learning through a hybrid supervised and alignment loss. We evaluate HWM against multiple model merging baselines across four tasks and four domains per task under diverse evaluation metrics. Experimental results demonstrate the effectiveness of HWM by showing that the generalist can be performant simultaneously in multi-domain multi-task learning. In future work, we plan to extend HWM to multimodal architectures and broader task settings.

6 Limitations

Limitations. Our work has several limitations that suggest directions for future research. First, we compare HWM with a range of representative model merging methods with publicly available implementations. However, several recent approaches do not release code, preventing direct empirical comparison. We also exclude methods designed only for merging two models or methods requiring additional supervision signals, such as user preferences, to ensure a fair comparison under our multi-model unsupervised merging setting.

Second, due to the cost of LLM-based evaluation, we apply the GPT-4.1 LLM-as-a-judge protocol to a randomly sampled subset of 1k test instances per domain and task. This results in 16k judged examples, covering approximately 8% of the full test set of over 191k instances. While this provides a substantial human-aligned evaluation, sampling may introduce variance and reduce statistical confidence compared with full-test evaluation. To mitigate this limitation, we evaluate all models on the full test set using automatic metrics, including BERTScore, SARI, METEOR, prediction validity, and Micro-F1, and observe trends consistent with the GPT-4.1 judgments.

Third, our experiments are conducted with a single backbone, LLaMA-3.1-8B-Instruct. Fixing the backbone allows us to isolate the effect of different merging strategies under controlled conditions, but it leaves open how well HWM generalizes to other model families and scales. Extending HWM to more backbones and incorporating additional recent merging methods are important directions for future work.

References

Martial Agueh and Guillaume Carlier. 2011. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924.

Takuya Akiba, Makoto Shing, Yujin Tang, Qi Sun, and David Ha. 2025. Evolutionary optimization of model merging recipes. *Nature Machine Intelligence*, 7(2):195–204.

Jason M Altschuler and Enric Boix-Adsera. 2021. Wasserstein barycenters can be computed in polynomial time in fixed dimension. *Journal of Machine Learning Research*, 22(44):1–19.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Miguel Romero Calvo, Shuoyang Ding, Corey D Barrett, Georgiana Dinu, and George Karypis. 2025. Beyond instruction-conditioning, MoTE: Mixture of task experts for multi-task embedding models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22731–22746, Vienna, Austria. Association for Computational Linguistics.

Ronald Cardenas, Bingsheng Yao, Dakuo Wang, and Yufang Hou. 2023. 'don't get too technical with me': A discourse structure-based framework for automatic science journalism. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1186–1202, Singapore. Association for Computational Linguistics.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multitask learning in natural language processing: An overview. *ACM Computing Surveys*, 56(12):1–32.

De Cheng, Zhipeng Xu, Xinyang Jiang, Nannan Wang, Dongsheng Li, and Xinbo Gao. 2024. Disentangled prompt representation for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23595–23604.

Ming Cheng, Jiaying Gong, and Hoda Eldardiry. 2025a. Sci-LoRA: Mixture of scientific LoRAs for cross-domain lay paraphrasing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18524–18541, Vienna, Austria. Association for Computational Linguistics.

Ming Cheng, Jiaying Gong, Chenhan Yuan, William A Ingram, Edward Fox, and Hoda Eldardiry. 2025b. VTechAGP: An academic-to-general-audience text paraphrase dataset and benchmark models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6110–6130, Albuquerque, New Mexico. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,*

Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Nicholas E. Corrado, Julian Katz-Samuels, Adithya M Devraj, Hyokun Yun, Chao Zhang, Yi Xu, Yi Pan, Bing Yin, and Trishul Chilimbi. 2025. AutoMix-Align: Adaptive data mixing for multi-task preference optimization in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20234–20258, Vienna, Austria. Association for Computational Linguistics.

MohammadReza Davari and Eugene Belilovsky. 2024. Model breadcrumbs: Scaling multi-task model merging with sparse masks. In *European Conference on Computer Vision*, pages 270–287. Springer.

Pala Tej Deep, Rishabh Bhardwaj, and Soujanya Poria. 2024. Della-merging: Reducing interference in model merging through magnitude-based sampling. *arXiv preprint arXiv:2406.11617*.

Hongxin Ding, Yue Fang, Runchuan Zhu, Xinke Jiang, Jinyang Zhang, Yongxin Xu, Weibin Liao, Xu Chu, Junfeng Zhao, and Yasha Wang. 2025. 3DS: Medical domain adaptation of LLMs via decomposed difficulty-based data selection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19473–19495, Suzhou, China. Association for Computational Linguistics.

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

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

Zitao Fang, Guodong Du, Shuyang Yu, Yifei Guo, Yiwei Zhang, Yiyao Cao, Jing Li, Ho-Kin Tang, and Sim Kuan Goh. 2025. To see a world in a spark of neuron: Disentangling multi-task interference for training-free model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15731–15751, Suzhou, China. Association for Computational Linguistics.

Zichuan Fu, Xian Wu, Yejing Wang, Wanyu Wang, Shanshan Ye, Hongzhi Yin, Yi Chang, Yefeng Zheng, and Xiangyu Zhao. 2025. Training-free LLM merging for multi-task learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 33111–33124, Vienna, Austria. Association for Computational Linguistics.

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics.

Tomas Goldsack, Zhihao Zhang, Chenghua Lin, and Carolina Scarton. 2022. Making science simple: Corpora for the lay summarisation of scientific literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10589–10604, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zi Gong, Hang Yu, Cong Liao, Bingchang Liu, Chaoyu Chen, and Jianguo Li. 2024. CoBa: Convergence balancer for multitask finetuning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8063–8077, Miami, Florida, USA. Association for Computational Linguistics.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. 2024. IEPile: Unearthing large scale schema-conditioned information extraction corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 127–146, Bangkok, Thailand. Association for Computational Linguistics.

Yue Guo, Wei Qiu, Gondy Leroy, Sheng Wang, and Trevor Cohen. 2024. Retrieval augmentation of large language models for lay language generation. *Journal of Biomedical Informatics*, 149:104580.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing: Student Research Workshop, pages 292–303, Online. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, Wei Zhou, and Songlin Hu. 2025. Impartial multi-task representation learning via variance-invariant probabilistic decoding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19883–19897, Vienna, Austria. Association for Computational Linguistics.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1419–1436, Online. Association for Computational Linguistics.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

Moritz Imfeld, Jacopo Graldi, Marco Giordano, Thomas Hofmann, Sotiris Anagnostidis, and Sidak Pal Singh. 2024. Transformer fusion with optimal transport. In *The Twelfth International Conference on Learning Representations*.

Dong-Hwan Jang, Sangdoo Yun, and Dongyoon Han. 2024. Model stock: All we need is just a few fine-tuned models. In *European Conference on Computer Vision*, pages 207–223. Springer.

Jiyue Jiang, Alfred Kar Yin Truong, Yanyu Chen, Qinghang Bao, Sheng Wang, Pengan Chen, Jiuming Wang, Lingpeng Kong, Yu Li, and Chuan Wu. 2025. Developing and utilizing a large-scale Cantonese dataset for multi-tasking in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1924–1944, Suzhou, China. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Birgit Kirsch, Héctor Allende-Cid, and Stefan Rueping. 2025. PM3-KIE: A probabilistic multi-task metamodel for document key information extraction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20890–20912, Vienna, Austria. Association for Computational Linguistics.

Yilun Kong, Guozheng Ma, Qi Zhao, Haoyu Wang, Li Shen, Xueqian Wang, and Dacheng Tao. 2025. Mastering massive multi-task reinforcement learning via mixture-of-expert decision transformer. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 31379–31396. PMLR.

Anastassia Kornilova and Vladimir Eidelman. 2019. BillSum: A corpus for automatic summarization of US legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 48–56, Hong Kong, China. Association for Computational Linguistics.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6536–6558, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wai-Chung Kwan, Xingshan Zeng, Yufei Wang, Yusen Sun, Liangyou Li, Yuxin Jiang, Lifeng Shang, Qun Liu, and Kam-Fai Wong. 2024. M4le: a multi-ability multi-range multi-task multi-domain long-context evaluation benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15568–15592.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

François Ledoyen, Gaël Dias, Jeremie Pantin, Alexis Lechervy, Fabrice Maurel, and Youssef Chahir. 2025. Facilitating cognitive accessibility with LLMs: A multi-task approach to easy-to-read text generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11782–11808, Suzhou, China. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, and 1 others. 2025a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791.

Erchao Li, Hongxu Li, and Yu Peng. 2025b. Competitive-cooperative multi-task optimization algorithm with historical success archive. *ACM Transactions on Knowledge Discovery from Data*, (0):000–000.

Zijian Li, Xiaocheng Feng, Huixin Liu, Yichong Huang, Ting Liu, and Bing Qin. 2025c. FroM: Frobenius norm-based data-free adaptive model merging. In

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 4674–4687, Suzhou, China. Association for Computational Linguistics.

Aoming Liu, Kevin Miller, Venkatesh Saligrama, Kate Saenko, Boqing Gong, Ser-Nam Lim, and Bryan A. Plummer. 2025a. Scaling up temporal domain generalization via temporal experts averaging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28196–28219, Suzhou, China. Association for Computational Linguistics.

Jinda Liu, Yi Chang, and Yuan Wu. 2025b. R-LoRA: Randomized multi-head LoRA for efficient multi-task learning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 660–674, Suzhou, China. Association for Computational Linguistics.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114.

Shuaiqi LIU, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Generating a structured summary of numerous academic papers: Dataset and method. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4259–4265. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Hongzhi Qi, Nan Bai, Jianqiang Li, Wei Zhai, Qing Zhao, Qi Gao, Bing Xiang Yang, and Guanghui Fu. 2025. Generalizable cross-lingual cognitive distortion detection with standardized annotations and multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16041–16051, Vienna, Austria. Association for Computational Linguistics.

Shanbao Qiao, Xuebing Liu, Akshat Gupta, and Seung-Hoon Na. 2025. SeqMMR: Sequential model merging and LLM routing for enhanced batched sequential knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16932–16947, Vienna, Austria. Association for Computational Linguistics.

Haiquan Qiu, You Wu, Dong Li, Jianmin Guo, and Quanming Yao. 2025. Superpose task-specific features for model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4200–4214, Suzhou, China. Association for Computational Linguistics.

Huaizhi Qu, Xinyu Zhao, Jie Peng, Kwonjoon Lee, Behzad Dariush, and Tianlong Chen. 2025. UQ-Merge: Uncertainty guided multimodal large language model merging. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1401–1417, Vienna, Austria. Association for Computational Linguistics.

Abudurexiti Reheman, Hongyu Liu, Junhao Ruan, Abudukeyumu Abudula, Yingfeng Luo, Tong Xiao, and JingBo Zhu. 2025. Enhancing neural machine translation through target language data: A kNN-LM approach for domain adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10053–10065, Vienna, Austria. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIG-PATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Sidak Pal Singh and Martin Jaggi. 2020. Model fusion via optimal transport. *Advances in Neural Information Processing Systems*, 33:22045–22055.

Jun Sun, Xinxin Zhang, Simin Hong, Jian Zhu, and Lingfang Zeng. 2025a. Adversarial alignment with anchor dragging drift (a^3d^2): Multimodal domain adaptation with partially shifted modalities. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19680–19690, Vienna, Austria. Association for Computational Linguistics.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangliao Geng, and Boyang Li. 2025b. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, New Orleans, Louisiana. Association for Computational Linguistics.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer

Suleman. 2017. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Alex Wang, Richard Yuanzhe Pang, Angelica Chen, Jason Phang, and Samuel R. Bowman. 2022. SQuALITY: Building a long-document summarization dataset the hard way. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1139–1156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Bowen Wang, Haiyuan Wan, Liwen Shi, Chen Yang, Peng He, Yue Ma, Haochen Han, Wenhao Li, Tiao Tan, Yongjian Li, Fangming Liu, Gong Yifan, and Sheng Zhang. 2025. RECALL: REpresentation-aligned catastrophic-forgetting ALLeviation via hierarchical model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16392–16406, Suzhou, China. Association for Computational Linguistics.

Mitchell Wortsman, Gabriel Ilharco, Samir Ya Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and 1 others. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *International conference on machine learning*, pages 23965–23998. PMLR.

Guofu Xie, Xiao Zhang, Ting Yao, and Yunsheng Shi. 2025. Bone soups: A seek-and-soup model merging approach for controllable multi-objective generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27237–27263, Vienna, Austria. Association for Computational Linguistics.

Yi Xin, Junlong Du, Qiang Wang, Ke Yan, and Shouhong Ding. 2024. Mmap: Multi-modal alignment prompt for cross-domain multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16076–16084.

Jiahui Xu, Lu Sun, and Dengji Zhao. 2024. Mome: Mixture-of-masked-experts for efficient multi-task recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2527–2531.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yangyang Xu, Xi Ye, and Duo Su. 2025. Multi-task dense prediction fine-tuning with mixture of fine-grained experts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4758–4767.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems*, 36:7093–7115.

Enneng Yang, Zhenyi Wang, Li Shen, Shiwei Liu, Guibing Guo, Xingwei Wang, and Dacheng Tao. 2024. Adamerging: Adaptive model merging for multi-task learning. In *The Twelfth International Conference on Learning Representations*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Yongxin Yang and Timothy M. Hospedales. 2015. A unified perspective on multi-domain and multi-task learning. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R. Fabbri, Irene Li, Dan Friedman, and Dragomir R. Radev. 2019. Scisummnet: a large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

En Yu, Jie Lu, Xiaoyu Yang, Guangquan Zhang, and Zhen Fang. 2025. Learning robust spectral dynamics for temporal domain generalization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. Language models are super mario: Absorbing abilities from homologous models as a free lunch. In *Forty-first International Conference on Machine Learning*.

Dacao Zhang, Kun Zhang, Shimao Chu, Le Wu, Xin Li, and Si Wei. 2025a. MoRE: A mixture of low-rank experts for adaptive multi-task learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1311–1324, Vienna, Austria. Association for Computational Linguistics.

Haobo Zhang and Jiayu Zhou. 2025. Unraveling LoRA interference: Orthogonal subspaces for robust model merging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26459–26472, Vienna, Austria. Association for Computational Linguistics.

Liang Zhang, Ziyao Lu, Fandong Meng, Hui Li, Jie Zhou, and Jinsong Su. 2025b. Advancing SMoE

for continuous domain adaptation of MLLMs: Adaptive router and domain-specific loss. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26584–26602, Vienna, Austria. Association for Computational Linguistics.

Peiyi Zhang, Richong Zhang, Zhijie Nie, and Ziqiao Wang. 2025c. Dynamic task vector grouping for efficient multi-task prompt tuning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26805–26821, Vienna, Austria. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhen Zhang, Yifan Yang, Kai Zhen, Nathan Susanj, Athanasios Mouchtaris, Siegfried Kunzmann, and Zheng Zhang. 2025d. MaZO: Masked zeroth-order optimization for multi-task fine-tuning of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18554, Suzhou, China. Association for Computational Linguistics.

Zijian Zhang, Shuchang Liu, Jiaao Yu, Qingpeng Cai, Xiangyu Zhao, Chunxu Zhang, Ziru Liu, Qidong Liu, Hongwei Zhao, Lantao Hu, and 1 others. 2024. M3oe: Multi-domain multi-task mixture-of experts recommendation framework. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 893–902.

Guangzhen Zhao, Yu Yao, Dechang Kong, and Zhenjiang Dong. 2025a. Seeking rational demonstrations for large language models: A domain generalization approach to unsupervised cross-domain keyphrase generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–424, Vienna, Austria. Association for Computational Linguistics.

Lulu Zhao, Weihao Zeng, Shi Xiaofeng, and Hua Zhou. 2025b. Mosld: An extremely parameter-efficient mixture-of-shared loras for multi-task learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1647–1659.

Ziyu Zhao, Yixiao Zhou, Zhi Zhang, Didi Zhu, Tao Shen, Zexi Li, Jinluan Yang, Xuwu Wang, Jing Su, Kun Kuang, and 1 others. 2025c. Each rank could be an expert: Single-ranked mixture of experts lora for multi-task learning. *arXiv preprint arXiv:2501.15103*.

Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *The World Wide Web Conference*, WWW '19, page 2472–2482, New York, NY, USA. Association for Computing Machinery.

Xun Zhu, Ying Hu, Fanbin Mo, Miao Li, and Ji Wu. 2024. Uni-med: a unified medical generalist foundation model for multi-task learning via connector-moe. *Advances in Neural Information Processing Systems*, 37:81225–81256.

A Appendix

A.1 Coordinate consistency of the shared support.

HWM assumes that the specialists share a common representation coordinate system, which holds in our setting because all specialists are fine-tuned from the same pretrained backbone and have the same architecture. The shared support $C = \{c_1, \ldots, c_K\}$ is not constructed separately for each specialist; instead, we pool representations from all specialists and fit K-means once on this union. This yields a single reference partition of the shared representation space. Moreover, the subsequent soft histograms and Wasserstein costs depend only on Euclidean distances to the shared centroids. Therefore, a global relabeling or permutation of representation coordinates applied consistently to all specialists would leave the distances, cluster assignments, and optimal transport costs unchanged. In contrast, arbitrary model-specific coordinate permutations would require an additional alignment step. Such permutations are unlikely in our experimental setup because all specialists are initialized from the same backbone rather than trained independently from scratch.

A.2 Dataset Comparisons

Table 1 summarizes the datasets used in our experiments across four NLP tasks and multiple domains. We construct a diverse evaluation suite covering paraphrasing, question answering, summarization, and information extraction. For each task, we include datasets from several domains, such as biomedical, news, science, Wikipedia, government, literature, and web sources. This design allows us to evaluate whether model merging methods can preserve both task-specific and domain-specific capabilities.

For paraphrasing, we include datasets from biomedical, news, science, and Wikipedia domains. For question answering, we cover web, biomedical, Wikipedia, and news domains. For summarization, we use datasets from news, literature, government, and science, including both single-document and multi-document summarization benchmarks. For information extraction, we use IEPile and group its constituent datasets into news, biomedical, science, and Wikipedia domains. Overall, the benchmark provides broad task and domain coverage, enabling a comprehensive evaluation of cross-domain model merging.

A.3 LLM-as-a-judge Instruction

We use the following four prompt templates for LLM-as-a-judge evaluation using GPT4.1 for task of text summarization, lay paraphrasing, generative question answering and information extraction, respectively.

Text Summarization System Prompt
You are an impartial grader of **TEXT SUMMARIZATION** quality.
Given a **SOURCE** document, a **CANDIDATE** summary, and (optionally) **REFERENCE** summaries, return a JSON object with score, verdict, and rationale.
Evaluation Dimensions (priority order):

- **Faithfulness**: all claims supported by SOURCE.
- **Coverage**: salient points from SOURCE.
- **Coherence & Fluency**: readable and grammatical.
- **Conciseness**: avoids unnecessary repetition.

Scoring ($[0, 1]$):

- **1.0**: fully faithful; strong coverage of key points; coherent and concise.
- **0.8–0.9**: faithful with minor omissions/redundancy; generally coherent.
- **0.6–0.7**: mostly faithful with omissions OR some verbosity/awkwardness.
- **0.3–0.5**: partial coverage with minor coherence issues.
- **0.1–0.2**: weak coverage.
- **0.0**: largely hallucinated or irrelevant.

Rules:

- Judge **only** against SOURCE.
- Use REFERENCES **only** for salience.
- Do not penalize paraphrasing.
- Penalize unsupported claims.
- Verbatim repetition is allowed if faithful.

Output (**STRICT JSON ONLY**):

```
{"score":[0,1],
 "verdict":"Correct|Partial|Incorrect",
 "rationale":"<=25 words"}
```

User Prompt
SOURCE DOCUMENT: {question}
REFERENCE SUMMARIES (optional): {references_block}
CANDIDATE SUMMARY: {candidate}

Lay Paraphrasing System Prompt
You are an impartial grader of **PARAPHRASING** quality.
Given a **SOURCE** sentence, a **CANDIDATE** paraphrase, and (optionally) **REFERENCE** paraphrases, return a JSON object with score, verdict, and rationale.
Evaluation Dimensions:

- **Meaning Preservation**: semantic equivalence to SOURCE.
- **Fluency & Grammaticality**: natural and correct language.
- **Paraphrase Diversity**: non-trivial rewording; penalize near-duplicates.
- **Factual Discipline**: no hallucinated or contradictory claims.

Scoring ($[0, 1]$):

- **1.0**: faithful, fluent, clearly reworded.
- **0.8–0.9**: faithful and fluent; minor stylistic issues.
- **0.6–0.7**: mostly faithful but close wording or minor omissions.
- **0.3–0.5**: partially faithful.
- **0.1–0.2**: marginal; poor faithfulness or fluency.
- **0.0**: incorrect meaning or hallucinated content.

Rules:

- Judge **semantic equivalence**, not surface similarity.
- Ignore harmless stylistic variation.
- Penalize contradictions or added information.

Output (**STRICT JSON ONLY**):

```
{"score":[0,1],
 "verdict":"Correct|Partial|Incorrect",
 "rationale":"<=25 words"}
```

User Prompt
SOURCE: {source}
REFERENCE PARAPHRASES (optional): {references_block}
CANDIDATE PARAPHRASE: {candidate}

Question Answering System Prompt
You are an impartial grader of **QUESTION ANSWERING** quality.
Given a **QUESTION**, a **CANDIDATE** answer, and one or more **REFERENCE** answers, return a JSON object with score, verdict, and rationale.
Evaluation Criteria:

- **Correctness**: semantically matches any reference answer.
- **Completeness**: includes required key details.
- **Consistency**: no contradictions or unsupported claims.

Scoring ($[0, 1]$):

- **1.0**: fully correct with no contradictions.
- **0.8–0.9**: almost correct, very minor omissions or phrasing differences.
- **0.6–0.7**: mostly correct, contains the main idea but lacks details.
- **0.3–0.5**: somewhat relevant with minor errors.
- **0.1–0.2**: marginally related but mostly incorrect.
- **0.0**: completely incorrect or irrelevant.

Rules:

- Judge **semantic equivalence**, not exact wording.
- Ignore harmless extra details; penalize unsupported/contradictory claims.
- If references differ, accept any correct alternative consistent with common knowledge.

Output (**STRICT JSON ONLY**):

```
{"score":[0,1],
 "verdict":"Correct|Partial|Incorrect",
 "rationale":"<=25 words"}
```

User Prompt
QUESTION: {question}
REFERENCE ANSWERS (one per line): {references_block}
CANDIDATE ANSWER: {candidate}

Information Extraction System Prompt
You are an impartial grader for **INFORMATION EXTRACTION** (event extraction).
Inputs:

- **SOURCE** text.
- **CANDIDATE_JSON**: possibly malformed JSON-like string {"event_type":[{"trigger":"...","arguments":{...}},...]}.
- **REFERENCE_JSONS**: one or more JSON-like strings with the same intent.

Procedure:

- **Parse & Repair**: recover best-effort JSON (accept single quotes, stray braces, repetitions).
- **Canonicalize**: lowercase; strip punctuation; collapse whitespace; lists de-duplicated (order irrelevant); nan/n/a/"" → NAN; **exact match only** after canonicalization.
- **Event Matching**: per event type, match events **1-to-1 by trigger**; unmatched candidate → FP_events; unmatched reference → FN_events.
- **Argument Scoring** (matched events only): ref=NAN ⇒ candidate value → FP_args; ref string/list must be present ⇒ TP_args, else FN_args; candidate extras → FP_args.

If both reference and candidate contain no events, set score=1.0.
Output (**STRICT JSON ONLY**):

```
{"score":[0,1],"precision":[0,1],"recall":[0,1],"f1":[0,1],
 "verdict":"Correct|Partial|Incorrect",
 "diagnostics":{"tp_events":0,"fp_events":0,"fn_events":0,
                "tp_args":0,"fp_args":0,"fn_args":0},
 "rationale":"<=25 words"}
```

Rules: no fuzzy matching; ignore harmless formatting; unrecoverable JSON ⇒ score=0.
User Prompt
SOURCE: {question}
REFERENCE_JSONS: {references_block}
CANDIDATE_JSON: {candidate}

Table 7: Domain-level ablation results for the hierarchical architecture. BS denotes BERTScore F1, Valid. denotes prediction validity rate, MiF1 denotes Micro-F1, and LLM denotes LLM-as-a-judge. Δ denotes HWM minus the global architecture.

Task	Metric	Domain	Global	HWM	Δ
Summ.	BERTScore	Gov.	78.85	79.94	+1.09
		Lit.	74.83	76.53	+1.70
		News	75.65	77.30	+1.65
		Sci.	70.58	77.75	+7.17
	SARI	Gov.	34.39	41.15	+6.76
		Lit.	39.86	39.54	-0.32
		News	37.86	42.09	+4.23
		Sci.	36.00	39.61	+3.61
	LLM	Gov.	57.97	67.28	+9.31
		Lit.	50.38	53.32	+2.94
		News	54.60	60.79	+6.19
		Sci.	66.30	78.40	+12.10
Lay	BERTScore	Bio.	79.67	81.74	+2.07
		News	81.59	79.27	-2.32
		Sci.	80.27	80.65	+0.38
		Wiki	82.01	83.33	+1.32
	SARI	Bio.	40.38	40.19	-0.19
		News	33.62	38.92	+5.30
		Sci.	32.39	34.52	+2.13
		Wiki	31.45	36.87	+5.42
	LLM	Bio.	65.52	70.14	+4.62
		News	39.43	43.49	+4.06
		Sci.	43.65	43.71	+0.06
		Wiki	38.42	44.77	+6.35
QA	BERTScore	Bio.	74.02	77.82	+3.80
		News	76.53	77.42	+0.89
		Web	76.96	77.80	+0.84
		Wiki	77.73	78.30	+0.57
	METEOR	Bio.	21.18	21.23	+0.05
		News	22.60	28.21	+5.61
		Web	23.74	28.45	+4.71
		Wiki	28.92	27.11	-1.81
	LLM	Bio.	66.92	67.17	+0.25
		News	42.57	45.40	+2.83
		Web	61.09	62.62	+1.53
		Wiki	71.88	79.69	+7.81
IE	Valid.	Bio.	67.96	85.71	+17.75
		News	85.35	98.25	+12.90
		Sci.	72.06	94.10	+22.04
		Wiki	80.13	94.12	+13.99
	MiF1	Bio.	6.16	12.34	+6.18
		News	28.67	36.60	+7.93
		Sci.	5.75	15.27	+9.52
		Wiki	22.92	37.14	+14.22
	LLM	Bio.	13.53	22.90	+9.37
		News	29.50	44.87	+15.37
		Sci.	11.88	20.48	+8.60
		Wiki	34.50	69.27	+34.77

Table 8: Domain-level ablation results for the number of cluster centroids. BS denotes BERTScore F1, Valid. denotes prediction validity rate, MiF1 denotes Micro-F1, and LLM denotes LLM-as-a-judge.

Task	Metric	Domain	HWM			HWM*		
			$K = 128$	$K = 256$	$K = 512$	$K = 128$	$K = 256$	$K = 512$
Summ.	BS	Gov.	77.00	77.84	79.94	75.83	75.83	77.28
		Lit.	74.15	74.85	76.53	73.42	73.40	77.41
		News	75.26	77.07	77.30	72.74	76.95	77.37
		Sci.	74.72	76.54	77.75	70.58	76.07	79.48
	SARI	Gov.	38.11	38.20	41.15	34.33	34.67	39.48
		Lit.	37.79	38.18	39.54	36.43	36.48	39.96
		News	40.35	42.96	42.09	39.50	42.75	42.24
		Sci.	37.73	39.36	39.61	36.00	39.27	40.21
	LLM	Gov.	56.87	67.58	67.28	56.49	57.08	63.00
		Lit.	30.84	36.65	53.32	50.59	50.59	53.23
		News	54.85	54.77	60.79	59.70	61.73	66.13
		Sci.	62.53	69.32	78.40	75.94	78.07	80.90
Lay	BS	Bio.	80.27	80.18	81.74	82.27	82.58	83.48
		News	77.39	79.58	79.27	79.24	79.52	79.84
		Sci.	80.52	80.32	80.65	79.24	79.56	79.12
		Wiki	83.20	83.46	83.33	74.07	74.50	77.70
	SARI	Bio.	40.07	40.22	40.19	42.80	42.97	42.48
		News	33.60	33.43	38.92	40.30	40.82	40.97
		Sci.	30.63	30.89	34.52	30.30	31.67	32.39
		Wiki	31.68	31.14	36.87	28.00	30.46	33.93
	LLM	Bio.	70.24	70.35	70.14	62.80	63.10	64.37
		News	41.82	42.88	43.49	40.00	40.60	45.06
		Sci.	42.67	42.42	43.71	36.50	38.40	36.49
		Wiki	36.80	38.19	44.77	34.10	39.90	46.48
QA	BS	Bio.	72.77	72.62	77.82	74.14	74.20	77.64
		News	77.32	77.36	77.42	73.19	76.60	80.39
		Web	76.49	77.10	77.80	80.27	80.15	79.54
		Wiki	77.47	78.42	78.30	79.30	79.98	78.00
	METEOR	Bio.	20.77	21.02	21.23	74.96	75.86	19.93
		News	21.92	21.46	28.21	22.21	20.57	31.50
		Web	27.05	27.55	28.45	27.50	27.46	30.91
		Wiki	26.99	26.99	27.11	24.90	27.56	27.65
	LLM	Bio.	65.42	65.83	67.17	60.30	63.47	65.13
		News	42.66	41.89	45.40	45.50	46.73	49.51
		Web	56.96	52.07	62.62	58.02	66.53	67.19
		Wiki	76.84	79.74	79.69	76.67	79.74	81.18
IE	Valid.	Bio.	82.73	80.66	85.71	82.03	84.00	85.66
		News	85.45	97.40	98.25	97.55	98.75	98.90
		Sci.	93.97	94.11	94.10	90.35	90.13	92.20
		Wiki	94.84	96.19	94.12	91.98	97.10	95.56
	MiF1	Bio.	10.57	7.09	12.34	10.09	10.24	11.81
		News	26.73	32.07	36.60	34.02	39.59	41.32
		Sci.	13.77	9.39	15.27	13.06	14.20	15.38
		Wiki	28.71	31.49	37.14	36.97	39.21	39.66
	LLM	Bio.	12.24	18.83	22.90	24.26	23.36	23.65
		News	40.91	42.10	44.87	44.98	44.80	47.22
		Sci.	14.18	18.18	20.48	20.57	20.38	22.98
		Wiki	69.16	64.36	69.27	68.69	63.19	70.39

Table 9: Domain-level ablation results for probing-set size. BS denotes BERTScore F1, Valid. denotes prediction validity rate, MiF1 denotes Micro-F1, and LLM denotes LLM-as-a-judge. Δ denotes the full probing setup minus the corresponding small probing setup.

Task	Metric	Domain	HWM			HWM*		
			Small	Full	Δ	Small	Full	Δ
Summ.	BS	Gov.	77.87	79.94	+2.07	75.80	77.28	+1.48
		Lit.	73.36	76.53	+3.17	73.28	77.41	+4.13
		News	74.61	77.30	+2.69	77.02	77.37	+0.35
		Sci.	74.66	77.75	+3.09	77.40	79.48	+2.08
	SARI	Gov.	38.45	41.15	+2.70	34.99	39.48	+4.49
		Lit.	37.73	39.54	+1.81	36.48	39.96	+3.48
		News	39.93	42.09	+2.16	42.99	42.24	-0.75
		Sci.	37.50	39.61	+2.11	39.38	40.21	+0.83
	LLM	Gov.	62.28	67.28	+5.00	65.10	63.00	-2.10
		Lit.	44.11	53.32	+9.21	49.95	53.23	+3.28
		News	58.42	60.79	+2.37	65.30	66.13	+0.83
		Sci.	77.12	78.40	+1.28	78.71	80.90	+2.19
Lay	BS	Bio.	79.52	81.74	+2.22	83.15	83.48	+0.33
		News	77.55	79.27	+1.72	80.57	79.84	-0.73
		Sci.	80.08	80.65	+0.57	73.40	79.12	+5.72
		Wiki	81.25	83.33	+2.08	81.88	77.70	-4.18
	SARI	Bio.	38.98	40.19	+1.21	44.01	42.48	-1.53
		News	33.87	38.92	+5.05	41.82	40.97	-0.85
		Sci.	30.59	34.52	+3.93	25.63	32.39	+6.76
		Wiki	33.75	36.87	+3.12	29.34	33.93	+4.59
	LLM	Bio.	66.43	70.14	+3.71	58.81	64.37	+5.56
		News	42.75	43.49	+0.74	43.71	45.06	+1.35
		Sci.	43.01	43.71	+0.70	36.19	36.49	+0.30
		Wiki	42.52	44.77	+2.25	44.95	46.48	+1.53
QA	BS	Bio.	74.25	77.82	+3.57	74.32	77.64	+3.32
		News	78.21	77.42	-0.79	80.52	80.39	-0.13
		Web	71.31	77.80	+6.49	80.11	79.54	-0.57
		Wiki	77.80	78.30	+0.50	78.91	78.00	-0.91
	METEOR	Bio.	22.25	21.23	-1.02	11.03	19.93	+8.90
		News	30.04	28.21	-1.83	31.36	31.50	+0.14
		Web	21.14	28.45	+7.31	27.16	30.91	+3.75
		Wiki	23.03	27.11	+4.08	25.99	27.65	+1.66
	LLM	Bio.	59.80	67.17	+7.37	63.76	65.13	+1.37
		News	40.40	45.40	+5.00	46.34	49.51	+3.17
		Web	56.04	62.62	+6.58	68.32	67.19	-1.13
		Wiki	70.20	79.69	+9.49	76.73	81.18	+4.45
IE	Valid.	Bio.	81.10	85.71	+4.61	84.57	85.66	+1.09
		News	94.15	98.25	+4.10	92.65	98.90	+6.25
		Sci.	89.29	94.10	+4.81	90.13	92.20	+2.07
		Wiki	84.53	94.12	+9.59	92.85	95.56	+2.71
	MiF1	Bio.	8.25	12.34	+4.09	10.86	11.81	+0.95
		News	32.10	36.60	+4.50	38.88	41.32	+2.44
		Sci.	9.91	15.27	+5.36	15.00	15.38	+0.38
		Wiki	33.60	37.14	+3.54	38.68	39.66	+0.98
	LLM	Bio.	23.86	22.90	-0.96	24.52	23.65	-0.87
		News	38.88	44.87	+5.99	46.95	47.22	+0.27
		Sci.	13.37	20.48	+7.11	18.61	22.98	+4.37
		Wiki	63.63	69.27	+5.64	68.32	70.39	+2.07