DELIFT: DATA EFFICIENT LANGUAGE MODEL INSTRUCTION FINE-TUNING

Ishika Agarwal[1], Krishnateja Killamsetty[2], Lucian Popa[2], Marina Danilevsky[2]
[1]University of Illinois Urbana-Champaign, [2]IBM Research
[1]ishikaa2@illinois.edu, [2]krishnateja.k@ibm.com, {lpopa, mdanile}@us.ibm.com

Year of Publication :- 2025

Introduction

- Fine-tuning Large Language Models (LLMs) is crucial for task specialization.
- It can become resource intensive due to redundant or uninformative data.
- Existing data selection methods rely on computationally expensive gradient-based metrics [2] or static embeddings that fail to adapt dynamically to the model's evolving state [3] that results in limiting their practical effectiveness.
- This paper proposes DELIFT (Data Efficient Language model Instruction Fine-Tuning).
- The method uses a novel, computationally efficient utility metric inspired by In-Context Learning (ICL)[1].

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
[2] Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). Less: Selecting influential data for targeted instruction tuning. arXiv preprint arXiv:2402.04333.
[3] Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., ... & Jin, H. (2023). Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701.

Introduction

- ICL-based metric measures the informational value of each data sample by quantifying its effectiveness as an in-context example in improving model predictions for other samples, reflecting its actual contribution relative to the model's current state.
- Submodular optimization methods introduced to systematically selects diverse, informative subsets optimized specifically for the following fine-tuning stage.
    - Instruction-tuning [4].
    - Task specific adaptation [5].
    - Continual fine tuning [6].

[4] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
[5] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
[6] Mazumder, S., & Liu, B. (2024). Continual Learning for Task-Oriented Dialogue Systems. In Lifelong and Continual Learning Dialogue Systems (pp. 127-151). Cham: Springer International Publishing.

Introduction

Primary contribution of this paper includes,

- A unified information-theoretic data selection paradigm.
- Submodular optimization framework.
- Extensive empirical validation.

Related Works

Data subset selection methods for Deep Neural Networks.

- Model-independent approaches.
  - Traditional model-independent techniques, such as clustering or distance metrics on pre-trained embeddings [7].
  - Does not get feedback from model, does not reflect model's changing state.
- Model-dependent approaches.
  - Model-dependent methods [8] incorporate the model's evolving knowledge by analyzing gradients or loss values.
  - However, performing gradient or influence estimations at scale becomes prohibitively expensive for large models.
- Subset selection with LLM feedback.
  - SelectIT [9] employs self-reflection prompts to rate data quality.
  - Filtering approaches [10] using GPT-4.
  - Though these provide a form of model-aware sampling, they typically lack a principled theoretical grounding.

In addition, all these approaches primarily target a single fine-tuning stage, limiting their adaptability for instruction tuning, task-specific adaptation, or continual learning.

[7] Bukharin, A., Li, S., Wang, Z., Yang, J., Yin, B., Li, X., ... & Jiang, H. (2023). Data diversity matters for robust instruction tuning. arXiv preprint arXiv:2311.14736.
[8] Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021, May). Glister: Generalization based data subset selection for efficient and robust learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 9, pp. 8110-8118).
[9] Liu, L., Liu, X., Wong, D. F., Li, D., Wang, Z., Hu, B., & Zhang, M. (2024). Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. arXiv preprint arXiv:2402.16705.
[10] Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., ... & Jin, H. (2023). Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701.

Methodology

- Goal is to identify a subset of data that maximizes the performance of large language models across the following three stages.
  - Instruction Tuning [4].
  - Task-Specific Fine-Tuning [5].
  - Continual Fine-Tuning [6].
- Two main parts of the approach.
  - Information-theoretic pairwise utility metric.
  - Submodular optimization to achieve data-efficient selection.
- Finally, it is shown how these components combine into solution for all fine-tuning stages.

[4] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
[5] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
[6] Mazumder, S., & Liu, B. (2024). Continual Learning for Task-Oriented Dialogue Systems. In Lifelong and Continual Learning Dialogue Systems (pp. 127-151). Cham: Springer International Publishing.

Pairwise Utility Metric

- Pairwise Utility Metric considering 2 training samples.

$$UF_{ij} = d(GT_i, \ p(y_i \mid x_i)) - d(GT_i, \ p(y_i \mid x_i, x_j, y_j))$$

- If d(·, ·) is chosen to be the Kullback-Leibler (KL) divergence, a simplified version would be,

$$UF_{ij} = \log \frac{p(y_i \mid x_i, x_j, y_j)}{p(y_i \mid x_i)} = \sum_{t=1}^{T} \log\left(\frac{p(y_{it} | x_i, x_j, y_j, y_{i,<t})}{p(y_{it} | x_i, y_{i,<t})}\right)$$

Pairwise Utility Metric

- In practice, for numerical stability, a length-normalized Euclidean distance rather than the KL-divergence was adopted,

$$d(GT_i, p(y_i \mid \cdot)) = \left\| 1 - p(y_i \mid \cdot) \right\|_2$$

- Computing UF_{ij} for all pairs (i, j) once before fine-tuning. Although this step is $O(n^2)$ in the dataset size, the cost is amortized because the same utility matrix can be reused for different fine-tuning stages or methods.

Submodular Optimization For Data Selection

- After computing UF_{ij} , a kernel matrix s_{ij} (e.g., set $s_{ij} = \max(UF_{ij}, 0)$) was defined.
- This kernel matrix was utilized in well studied submodular functions.
- Submodularity naturally captures diminishing returns and promotes coverage of diverse yet informative samples.
- Three submodular objectives:
    - Facility Location (FL) - Select a representative subset that covers the diversity of the whole dataset
    - Facility Location Mutual Information (FLMI) - Select a subset that maximizes mutual coverage between two sets.
    - Facility Location Conditional Gain (FLCG) - Select a subset A that adds new information beyond what another already-selected set Q provides.

Experimental Results

- Extensive experiments to evaluate DELIFT in three fine-tuning scenarios.
  - Instruction Tuning [4].
  - Task-Specific Fine-Tuning [5].
  - Continual Fine-Tuning [6].
- DELIFT was evaluated on variety of LLMs covering different parameter scales.
  - Base LLMs (Llama-3.2-3B, Mistral-7B-v0.1, opt-125m).
  - Instruction-tuned LLMs (Qwen2-72B-Instruct, Phi-3-mini-128k-instruct).
- Tests were done with different adaptation strategies ICL [1], QLoRA [11], and, for smaller models, full fine-tuning.

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
[4] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
[5] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
[6] Mazumder, S., & Liu, B. (2024). Continual Learning for Task-Oriented Dialogue Systems. In Lifelong and Continual Learning Dialogue Systems (pp. 127-151). Cham: Springer International Publishing.
[11] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115.

Datasets

- Instruction Tuning.
  - Mix-Instruct [12] .
  - P3 [13].
- Task-Specific Fine-Tuning.
  - HotpotQA [14] aligned with MMLU [15].
  - Mix-Instruct aligned with MT-Bench [16].
  - Mix-Instruct aligned with GSM-8k [17].
- Continual Fine-Tuning.
  - SQuAD [18] paired with HotpotQA to inject more complex, multi-hop reasoning data after simpler QA.
  - Proprietary IBM/Government domain query rewriting dataset.

Fixed an approximate budget of 30% for subset selection unless otherwise noted

[12] Jiang, D., Ren, X., & Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. arXiv preprint arXiv:2306.02561.
[13] Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207.
[15] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.
[16] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2009). Measuring massive multitask language understanding, 2021. URL https://arxiv. org/abs, 20.
[17] Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623.
[18] Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Baselines

- Full data baseline.
- Following other baseline data selection strategies.
  - Random - Random selection of 30 % of the data.
  - SelectIT [9] - Generates self-reflection prompts within the LLM to rate data quality, filtering out lower-quality samples.
  - LESS [2] - Employs gradient-based influence estimates, approximated via LoRA, to identify highly impactful data points for model parameter updates.
  - DEFT-UCS - Uses sentence embeddings to cluster the dataset for diversity; although it captures semantic variety, it lacks explicit model feedback to guide selection.
  - DELIFT (SE) - Operates the same submodular optimization as DELIFT but replaces our utility based kernel with static sentence embedding similarities.

[2] Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). Less: Selecting influential data for targeted instruction tuning. arXiv preprint arXiv:2402.04333.
[9] Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. arXiv preprint arXiv:2402.16705.

Metrics

- ROUGE [19] - Focuses on n-gram overlap for summarization tasks or generative text alignment.
- BGE [20] - Evaluates semantic similarity through the dot product of normalized sentence embeddings.
- LAJ [21] (LLM-as-Judge) - Assigns a 1–5 rating reflecting correctness, clarity, and instruction adherence.
- Classification Accuracy - Used primarily for multiple-choice tasks like MMLU.

[19] Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
[20] Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J. Y. (2024, July). C-pack: Packed resources for general chinese embeddings. In Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval (pp. 641-649).
[21] Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., ... & Seo, M. (2023, October). Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations.

USE CASE 1: INSTRUCTION TUNING

- Compare DELIFT with baselines on Mix-Instruct

Model	**Qwen2**						**Phi-3**					
Method	**ICL**			**QLoRA**			**ICL**			**QLoRA**		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	37.87	78.92	2.98	36.36	82.55	3.02	25.76	43.34	1.42	35.50	80.46	2.58
Random	39.00	80.66	3.12	44.45	85.46	3.12	33.05	72.73	2.92	44.70	83.75	2.95
SelectIT	43.08	84.50	3.18	45.14	85.88	3.21	36.11	76.31	3.18	49.68	85.84	3.20
LESS	42.08	83.24	3.26	45.16	84.95	3.28	47.10	**85.94**	3.23	48.68	**85.86**	3.24
DELIFT (SE)	47.43	84.40	3.28	48.22	86.50	3.28	46.62	85.28	3.24	45.64	83.70	3.27
DELIFT	**48.46**	**85.77**	**3.35**	**52.79**	**88.04**	**3.37**	**49.83**	85.27	**3.32**	**50.31**	84.40	**3.33**
Full Data	*58.65*	*88.72*	*3.45*	*65.51*	*92.24*	*3.51*	*55.92*	*88.26*	*3.45*	*74.98*	*93.33*	*3.84*

USE CASE 1: INSTRUCTION TUNING

- Compare DELIFT with baselines on P3.

Model	Qwen2						Phi-3					
Method	ICL			QLoRA			ICL			QLoRA		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	18.03	59.13	1.54	20.15	58.38	1.78	20.10	48.66	1.36	20.64	49.17	1.39
Random	20.05	59.39	1.79	20.29	59.39	1.83	20.83	49.92	2.24	24.51	53.41	2.36
SelectIT	31.38	71.08	2.86	32.96	74.76	2.90	35.37	66.67	2.52	38.98	69.84	2.54
LESS	34.59	83.23	3.07	35.03	83.37	3.50	39.69	72.12	3.17	40.32	70.89	3.24
DELIFT (SE)	34.69	83.31	3.43	35.46	83.43	3.53	37.07	71.49	3.52	38.13	79.68	3.74
DELIFT	**35.48**	**83.69**	**3.58**	**35.60**	**83.64**	**3.54**	**40.66**	**84.00**	**3.68**	**41.91**	**84.53**	**3.76**
Full Data	*36.43*	*84.25*	*3.53*	*35.88*	*76.87*	*3.63*	*42.07*	*85.26*	*3.78*	*44.73*	*87.03*	*3.82*

Evaluation of a base model

- Experimentation done on Llama-3.2-3B (base) which is a non-instruction tuned model

Method	**ICL**			**QLoRA**		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	25.88	61.97	1.90	28.51	73.14	2.48
Random	28.64	77.14	2.59	38.79	79.89	2.53
SelectIT	42.67	81.54	2.67	45.87	84.67	2.61
DEFT-UCS	41.55	80.12	2.63	41.03	81.86	2.59
LESS	44.99	82.48	2.69	50.54	84.14	2.78
DELIFT (SE)	51.19	83.54	2.72	55.32	85.92	2.79
DELIFT	**54.58**	**88.29**	**2.83**	**58.57**	**90.98**	**2.98**
Full Data	55.46	92.71	3.31	61.23	95.52	3.10

USE CASE 2: TASK-SPECIFIC FINE-TUNING

- Experiments done after getting subset from HotPotQA for MMLU task.

Method	Qwen2 (QLoRA)	Phi-3 (QLoRA)
Initial	82.10	69.10
Random	79.31	65.16
SelectIT	79.13	65.24
LESS	80.35	66.72
DELIFT (SE)	80.10	66.36
DELIFT	**81.70**	**68.70**
Full Data	*78.36*	*64.50*

USE CASE 2: TASK-SPECIFIC FINE-TUNING

- Experiments done after getting subset from Mix-Instruct for MT-Bench task

Model	Qwen2						Phi-3					
Method	ICL			QLoRA			ICL			QLoRA		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	44.32	74.86	2.48	47.65	77.92	2.72	39.57	69.43	2.31	42.89	72.76	2.53
Random	49.78	79.54	2.83	52.91	82.67	3.05	44.63	74.28	2.62	47.85	77.39	2.84
SelectIT	54.92	83.71	3.12	57.86	86.59	3.31	49.75	78.64	2.91	52.68	81.52	3.13
LESS	59.63	85.89	3.29	62.74	88.72	3.48	54.82	81.95	3.08	57.73	84.67	3.29
DELIFT (SE)	62.85	86.94	3.38	65.83	89.76	3.57	57.69	82.87	3.17	60.54	85.59	3.38
DELIFT	**64.73**	**87.82**	**3.47**	**67.91**	**90.64**	**3.66**	**59.58**	**83.76**	**3.26**	**62.47**	**86.48**	**3.47**
Full Data	*65.89*	*88.65*	*3.55*	*69.72*	*91.53*	*3.74*	*60.76*	*84.59*	*3.34*	*64.31*	*87.42*	*3.55*

FURTHER EXPERIMENTS ON COMPLEX REASONING TASKS

- Experiments done after getting subset from Mix-Instruct for GSM-8k.

Method	ICL			QLoRA		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	17.06	55.13	1.35	27.95	65.92	1.36
Random	19.18	56.03	1.38	33.85	81.74	1.85
SelectIT	30.27	77.23	1.45	42.29	88.17	2.49
DEFT-UCS	30.06	76.99	1.59	41.45	87.84	2.08
LESS	31.69	77.87	2.26	43.86	88.22	2.60
DELIFT (SE)	31.84	78.62	2.44	43.04	90.53	2.54
DELIFT (FL instead of FLMI)	32.30	78.54	2.54	44.62	91.04	2.63
DELIFT	**33.25**	**79.10**	**2.56**	**46.12**	**92.97**	**2.71**
Full Data	35.33	81.59	2.79	49.10	94.57	2.85

USE CASE 3: CONTINUAL FINE-TUNING

- IBM → GOVERNMENT

Model	Qwen2						Phi-3					
Method	ICL			QLoRA			ICL			QLoRA		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	44.11	70.49	2.43	48.49	80.85	2.62	40.66	58.68	1.52	43.96	69.56	2.29
Random	55.57	85.26	2.91	55.52	85.53	2.94	45.76	76.19	2.45	58.94	82.41	2.89
SelectIT	63.07	86.38	3.18	65.42	87.50	3.20	63.49	85.27	2.96	64.09	85.07	3.16
LESS	64.28	85.41	3.29	69.85	89.33	3.45	66.01	87.20	3.19	67.53	88.17	3.22
DELIFT (SE)	61.07	85.16	3.45	74.05	**92.47**	3.58	68.84	88.46	3.32	69.30	88.62	3.35
DELIFT	**69.49**	**87.94**	**3.60**	**74.19**	92.23	**3.65**	**74.11**	**89.41**	**3.57**	**74.38**	**91.55**	**3.57**
Full Data	*66.08*	*87.84*	*3.65*	*76.83*	*92.63*	*3.74*	*71.23*	*91.10*	*3.52*	*77.12*	*91.10*	*3.64*

USE CASE 3: CONTINUAL FINE-TUNING

- SQUAD → HotpotQA

Model	**Qwen2**						**Phi-3**					
Method	**ICL**			**QLoRA**			**ICL**			**QLoRA**		
	ROUGE	**BGE**	**LAJ**	**ROUGE**	**BGE**	**LAJ**	**ROUGE**	**BGE**	**LAJ**	**ROUGE**	**BGE**	**LAJ**
Initial	51.51	66.97	1.77	54.18	78.27	2.50	40.42	58.23	1.26	40.94	58.12	1.29
Random	54.38	79.12	2.57	59.23	82.02	2.66	44.29	59.45	1.33	50.29	61.52	1.60
SelectIT	58.03	83.75	2.82	63.26	84.01	2.87	47.35	74.15	2.54	56.88	80.47	2.70
LESS	67.16	85.76	2.94	69.72	86.63	3.26	60.97	81.41	2.84	61.56	81.53	2.88
DELIFT (SE)	73.75	88.01	3.26	74.84	88.79	3.30	64.44	83.95	3.03	66.35	84.77	3.14
DELIFT	**76.94**	**90.41**	**3.33**	**77.56**	**89.99**	**3.34**	**66.55**	**84.65**	**3.25**	**67.09**	**85.17**	**3.32**
Full Data	*77.78*	*90.31*	*3.35*	*78.72*	*90.77*	*3.48*	*68.47*	*85.93*	*3.33*	*70.48*	*86.06*	*3.44*

Ablation Studies

- Full Fine-Tuning vs. QLora

Method	QLoRA			Full Fine-Tuning		
	ROUGE	BGE	LAJ	ROUGE	BGE	LAJ
Initial	9.04	40.50	1.19	9.57	40.86	1.14
Random	12.55	46.99	1.25	13.07	47.91	1.30
SelectIT	14.78	49.80	1.26	15.35	50.42	1.28
DEFT-UCS	15.12	50.29	1.39	15.16	50.29	1.39
LESS	15.52	50.70	1.38	16.02	51.30	1.40
DELIFT (SE)	18.81	55.02	1.35	17.06	55.93	1.38
DELIFT	**19.72**	**57.98**	**1.43**	**19.87**	**58.11**	**1.45**
Full Data	*20.39*	*60.39*	*1.95*	*21.64*	*61.70*	*2.05*

Ablation Studies

- Comparing Submodular Objectives
  - Although FL alone can beat naive baselines in specialized or incremental settings, the specialized objectives (FLMI for domain tasks, FLCG for continual updates) yield stronger alignment.
  - This underscores the importance of matching the submodular objective to the fine-tuning stage.
- Ablation on Subset size
  - Experimentation was done on how varying the subset size influences performance.
  - Subset sizes ranging of 5 % to 50 % of the original dataset were tested.
  - Performance under all methods generally improves with larger subsets but exhibits diminishing returns beyond 30-40 %.
  - DELIFT consistently yields higher LAJ scores than other baselines at every subset size, demonstrating its robustness even under very aggressive pruning

Key Observations

- Utility-based kernel outperforms static or gradient-based methods.
- Stage-specific objectives (FL, FLMI, FLCG).
- Significant pruning (up to 70%).
- Method-agnostic gains.

Conclusion

- A novel DELIFT, a novel approach to data-efficient fine-tuning of large language models by employing a versatile pairwise utility metric combined with submodular optimization techniques for optimal data selection.
- Empirical evaluations showed that DELIFT can reduce data and computational requirements by up to 70% while achieving performance comparable to the full dataset, and outperforming existing data selection methods by up to 26% in effectiveness.
- Has a risk of bias amplification in the selected data.

References

1) Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
2) Xia, M., Malladi, S., Gururangan, S., Arora, S., & Chen, D. (2024). Less: Selecting influential data for targeted instruction tuning. arXiv preprint arXiv:2402.04333.
3) Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., ... & Jin, H. (2023). Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701.
4) Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
5) Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
6) Mazumder, S., & Liu, B. (2024). Continual Learning for Task-Oriented Dialogue Systems. In Lifelong and Continual Learning Dialogue Systems (pp. 127-151). Cham: Springer International Publishing.
7) Bukharin, A., Li, S., Wang, Z., Yang, J., Yin, B., Li, X., ... & Jiang, H. (2023). Data diversity matters for robust instruction tuning. arXiv preprint arXiv:2311.14736.
8) Killamsetty, K., Sivasubramanian, D., Ramakrishnan, G., & Iyer, R. (2021, May). Glister: Generalization based data subset selection for efficient and robust learning. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 9, pp. 8110-8118).
9) Liu, L., Liu, X., Wong, D. F., Li, D., Wang, Z., Hu, B., & Zhang, M. (2024). Selectit: Selective instruction tuning for large language models via uncertainty-aware self-reflection. arXiv preprint arXiv:2402.16705.
10) Chen, L., Li, S., Yan, J., Wang, H., Gunaratna, K., Yadav, V., ... & Jin, H. (2023). Alpagasus: Training a better alpaca with fewer data. arXiv preprint arXiv:2307.08701.
11) Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115.
12) Jiang, D., Ren, X., & Lin, B. Y. (2023). Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. arXiv preprint arXiv:2306.02561.
13) Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. arXiv preprint arXiv:2110.08207.
14) Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., & Manning, C. D. (2018). HotpotQA: A dataset for diverse, explainable multi-hop question answering. arXiv preprint arXiv:1809.09600.
15) Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2009). Measuring massive multitask language understanding, 2021. URL https://arxiv. org/abs, 20.
16) Zheng, L., Chiang, W. L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., ... & Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36, 46595-46623.
17) Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., ... & Schulman, J. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
18) Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.
19) Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).
20) Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., & Nie, J. Y. (2024, July). C-pack: Packed resources for general chinese embeddings. In Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval (pp. 641-649).
21) Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., ... & Seo, M. (2023, October). Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations.