Learning to Maximize Mutual Information for Chain-of-Thought Distillation

Xin Chen[1], **Hanxian Huang**[2], **Yanjun Gao**[3], **Yi Wang**[1], **Jishen Zhao**[2], **Ke Ding**[1]
[1]Applied ML Group, Intel Corp.
[2]University of California San Diego, [3]University of Wisconsin Madison
[1]{xin.chen, yi.a.wang, ke.ding}@intel.com,
[2]{hah008, jzhao}@ucsd.edu,
[3]ygao@medicine.wisc.edu

Abstract

Knowledge distillation, the technique of transferring knowledge from large, complex models to smaller ones, marks a pivotal step towards efficient AI deployment. Distilling Step-by-Step (DSS), a novel method utilizing chain-of-thought (CoT) distillation, has demonstrated promise by imbuing smaller models with the superior reasoning capabilities of their larger counterparts. In DSS, the distilled model acquires the ability to generate rationales and predict labels concurrently through a multi-task learning framework. However, DSS overlooks the intrinsic relationship between the two training tasks, leading to ineffective integration of CoT knowledge with the task of label prediction. To this end, we investigate the mutual relationship of the two tasks from Information Bottleneck perspective and formulate it as maximizing the mutual information of the representation features of the two tasks. We propose a variational approach to solve this optimization problem using a learning-based method. Our experimental results across four datasets demonstrate that our method outperforms the state-of-the-art DSS. Our findings offer insightful guidance for future research on language model distillation as well as applications involving CoT. Codes are available at https://github.com/xinchen9/cot_distillation_ACL2024

1 Introduction

The capabilities of larger language models (LLMs) tend to scale with their model size, leading to a substantial demand for memory and compute resources (Chowdhery et al., 2023; Wei et al., 2022a). Distilling knowledge from larger LLMs to smaller LLMs has been crucial for the efficient deployment of AI (Hinton et al., 2015; Phuong and Lampert, 2019). Chain-of-Thought (CoT) (Wei et al., 2022b) distillation represents a pivotal advance in the quest to endow smaller language models with the sophisticated reasoning capabilities of their larger counterparts. By distilling complex thought processes into more compact models, this approach aims to democratize access to advanced natural language understanding and reasoning across a wider array of computational resources (Ma et al., 2023; Magister et al., 2023; Li et al., 2023). .

Distilling Step-by-Step (DSS) (Hsieh et al., 2023) introduces a CoT distillation method that guides smaller models using rationales from LLMs within a multi-task learning (MTL) framework, training them for both *label prediction* and *rationale generation* tasks. This framework simultaneously optimizes the model for two related objectives on the same input, enhancing its chain-of-thought learning by sharing representations between the two tasks, thereby improving overall performance efficiently. While DSS brings out the benefits of reducing computational costs, it suffers from the same problem as the conventional MTL framework, that is the difficulty in effectively connecting the prediction and generation tasks. The intricacies inherent in training models within the MTL framework can undermine the effectiveness and reliability of the DSS process (Wang et al., 2023b). Despite the successful setup of an MTL framework in DSS, where the tasks of label prediction and rationale generation are intrinsically related, the current configuration may not optimally capture and maximize the mutual knowledge between these tasks. Furthermore, LLMs are prone to producing hallucinations and inconsistent rationales, which potentially mislead the student model toward incorrect answers and cause conflicts in MTL that destruct student model learning (Mueller et al., 2022).

To address this issue, we model the DSS using information bottleneck and investigate it from an information-theoretic viewpoint (Tishby and Zaslavsky, 2015). Subsequently, we formulate the DSS as an optimization problem to maximize mutual information (MI) of label prediction and ra-

Findings of the Association for Computational Linguistics: ACL 2024, pages 6857–6868

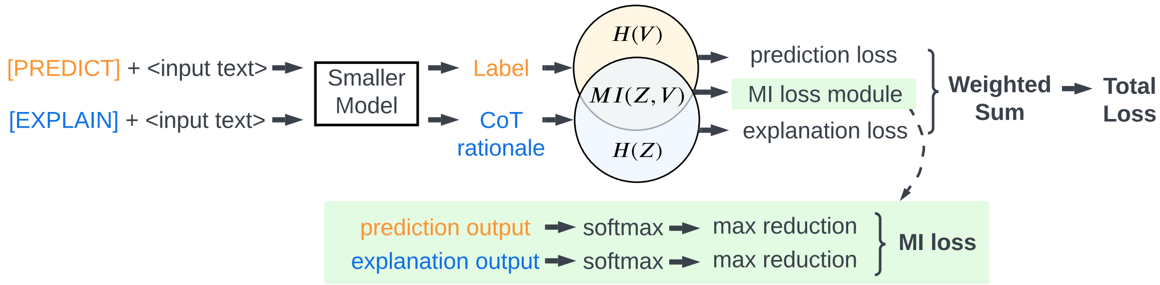

Figure 1: Overview of our approach: CoT distillation from an IB perspective and measurement of the intrinsic relationship between the two tasks by MI. The DSS is an MTL framework pipeline comprising label prediction and rationale generation tasks. H represents the entropy of representation features V and Z. Besides prediction loss and explanation losses used in conventional DSS, we design an auxiliary loss module to maximize MI between the two representation features. This process enhances CoT reasoning capacity through knowledge distillation.

tionale generation tasks. However, estimating MI from finite data has historically been a difficult problem in both deep learning and information theory (McAllester and Stratos, 2020; Belghazi et al., 2018; Paninski, 2003).

In this study, we introduce a variational method to estimate the MI. We propose a practical yet effective auxiliary loss to quantify the shared information between the prediction and the generation tasks, thereby enhancing the alignment between the two tasks and facilitating the knowledge transfer from CoT. We conduct comprehensive experiments with two smaller types of T5 models (Raffel et al., 2020), T5-base (220M) and T5-small (60M), on four popular datasets. Furthermore, we provide detailed analysis in Section 5. Our main contributions are summarized below:

- We reframe the MTL framework of DSS as a MI estimation challenge, aiming to maximize the MI between label prediction and rationale generation tasks. To achieve this, we introduce a variational approach grounded in the IB principle for effective MI estimation. To the best of our knowledge, we present the first work of improving CoT distillation from an IB perspective.
- Beyond establishing a theoretical foundation, we present a practical approach for MI estimation, incorporating a simple yet effective auxiliary loss to learning to maximize MI and enhance DSS.
- Our methodology demonstrably outperforms existing benchmarks across multiple datasets, evidencing the efficacy of our approach in enhancing the reasoning capabilities of distilled models.
- We conduct a systematic review of the relationship between label prediction and rationale generation tasks under MTL training, presenting both qualitative and quantitative analysis results.

Armed with theoretical proofs and experimental results, we aim to lay the groundwork for future research on enhancing CoT distillation within an effective MTL framework, guided by principles from information theory.

2 Related Work

We present an overview of previous work across three areas related to our study: knowledge distillation, multi-task learning, and information bottleneck.

Knowledge Distillation (KD) Originally designed to train small models by leveraging the extensive knowledge of larger models (Hinton et al., 2015), KD has since been extended to a variety of applications, owwwing to its effective transfer of knowledge across models and tasks (Chen et al., 2021; Wang and Yoon, 2021; Sanh et al., 2019; Jiao et al., 2020; Luo et al., 2024; Go et al., 2023; Zhang et al., 2022b). A crucial yet open challenge is how to effectively transfer the knowledge. To address the issue, previous studies (Zhang et al., 2022c; Allen-Zhu and Li, 2023; Zhang et al., 2021) have extracted various features and designed auxiliary loss functions to enhance KD. Our work focuses on improving the model by acquiring mutual knowledge to address both label prediction and rationale generation tasks.

Multi-Task Learning (MTL) By exploiting commonalities and differences among relevant tasks, MTL can enhance learning efficiency and improve prediction accuracy by learning multiple objectives from a shared representation (Caruana, 1997; Zhang and Yang, 2021). In recent years, MLT has been broadly applied to NLP tasks (Worsham and Kalita, 2020; Zhang et al., 2023b; Liu et al., 2019). However, some studies have identified that training multiple tasks trained simultaneously could lead to conflicts among them, making it challenging to optimize the performance of all tasks simultaneously (Kendall et al., 2018; Lin et al., 2019). Recently, KD has also been applied within MTL frameworks, achieving state-of-the-art results in various applications (Li and Bilen, 2020; Xu et al., 2023; Yang et al., 2022; Garner and Dux, 2023; Zhang et al., 2023a).

Information Bottleneck (IB) IB (Tishby and Zaslavsky, 2015; Slonim, 2002) provides a powerful statistical tool to learn representation to preserve complex intrinsic correlation structures over high dimensional data. As a general measure of the dependence between two random variables, MI is also widely used in deep learning to effectively represent feature dependencies (Cover, 1999; Covert et al., 2023; Liu et al., 2009). Estimating MI is known to be challenging, and recent progress has been made towards learning-based variational approaches (Tian et al., 2020; Covert et al., 2023; Bachman et al., 2019; Tschannen et al., 2019; Belghazi et al., 2018; Diao et al., 2023). Another challenge associated with the IB principle is the optimization process, which involves a trade-off between achieving concise representation and maintaining strong predictive capabilities (Alemi et al., 2016; Wang et al., 2019). Consequently, optimizing IB becomes a complex task that heavily depends on the problem formulation and the provision of an effective optimization solution. Recent studies have applied IB to solve complex machine learning problems both in computer vision (Tian et al., 2021; Du et al., 2020; Wan et al., 2021) and NLP (Chen and Ji, 2020; Zhang et al., 2022a; Paranjape et al., 2020). In this paper, we formulate our CoT distillation problem with MTL training pipeline using IB method, and provide a learning-based solution to optimize IB for our CoT distillation, as detailed in Section 3.

3 Methodology

This section begins with an introduction to preliminaries of IB. Following this, we formulate our CoT distillation idea within the IB framework and propose a learning approach to optimize MI.

3.1 Preliminaries

Under the DSS framework, the prefixes [PREDICT] and [EXPLAIN] will be prepended to the input text, TEXT, for tasks corresponding to label prediction and rationale generation, respectively. In the label prediction task, given the input [PREDICT] + TEXT along with predictive labels $\mathbf{Y}$, a representation feature $\mathbf{V}$, where $\mathbf{V} \in \mathbb{R}^d$, is trained using $\mathbf{Y}$. Similarly, in the rationale generation task, the input [EXPLAIN] + TEXT and rationale label $\mathbf{R}$ guide the training of a representation feature $\mathbf{Z}$, where $\mathbf{Z} \in \mathbb{R}^d$, using $\mathbf{R}$.

Our goal is to distill CoT knowledge from larger LLMs to smaller LLMs models. To achieve this, based on the basis of IB (Tishby and Zaslavsky, 2015; Zhang et al., 2022a; Wang et al., 2019), we model the DSS as following:

$$I(Z;Y) = \int p(z,y) \log \frac{p(z,y)}{p(z)p(y)} dzdy. \quad (1)$$

where sampling observations $z \sim \mathbf{Z}$ and $v \sim \mathbf{V}$. Here $p(\cdot)$ represents the probability distribution.

To encourage CoT distillation to focus on the information represented in label $\mathbf{Y}$, we propose using IB to enforce an upper, bound I_c, on the information flow from the representation features V to the representation features Z. This is achieved by maximizing the following objective:

$$\max I(Z;Y) \;\; s.t. \;\; I(Z;V) \leq I_c. \quad (2)$$

By employing Lagrangian objective, IB allows Z to be maximally expressive about Y while being maximally compressive regarding the input data, as follows:

$$\mathcal{L}_{IB} = I(Z;V) - \beta I(Z;Y) \quad (3)$$

where β is the Lagrange multiplier. Clearly, Eq. 3 demonstrates the trade-off optimization between high mutual information and high compression (Zhang et al., 2022a; Alemi et al., 2016). In our scenario, given a predefined small student model, the compression ratio is fixed. Therefore, we formulate the CoT distillation as an optimization problem:

$$\max I(Z;V) \quad (4)$$

Due to symmetric property of MI, $I(Z;V) = I(V;Z)$. CoT distillation can also enhance rationale generation task with the label knowledge. This is validated in Section 5.

3.2 Variational Bounds of MI

We rewrite $I(Z;V)$ of Equation 4 as follows:

$$I(Z;V) = \mathbb{E}_{p(z,v)}\Big[\log \frac{p(v|z)}{p(v)}\Big] \quad (5)$$

According to (Poole et al., 2019; Covert et al., 2023), a tractable variational upper bound can be established by introducing a variational approximation $q(v)$ to replace the intractable marginal $p(v)$, demonstrated by:

$$\begin{aligned} I(Z;V) &= \mathbb{E}_{p(z,v)}\Big[\log \frac{p(v|z)q(v)}{p(v)q(v)}\Big] \\ &= \mathbb{E}_{p(z,v)}\Big[\log \frac{p(v|z)}{q(v)}\Big] \\ &- KL(p(v)||q(v)) \end{aligned} \quad (6)$$

here $KL[\cdot||\cdot]$ denotes Kullback-Leibler divergence. The bound is tight when $q(v) = p(v)$. Consequently, $KL(p(v)||q(v))$ is equal to $KL(p(v)||p(v))$, which becomes zero. Therefore, we can derive at the following inequality:

$$I(Z;V) \leq \mathbb{E}_{p(z,v)}\Big[\log \frac{p(z|v)}{p(v)}\Big] \quad (7)$$

We can then express the MI from Eq. 5 as the follows:

$$\begin{aligned} \mathbb{E}_{p(z,v)}\Big[\log \frac{p(z|v)}{p(v)}\Big] &= \sum p(z,v) \log \frac{p(v|z)}{p(v)} \\ &= \sum p(z|v)p(v) \log p(v|z) \\ &- \sum p(v)p(z|v) \log p(v) \end{aligned} \quad (8)$$

Assuming that $p(v)$ is uniform distribution for maximal entropy (Schröder and Biemann, 2020), the term $\sum p(v)p(z|v)\log p(v)$ is considered as a constant. This also allows for the omission of $p(v)$ in $\sum p(z|v)p(v)\log p(v|z)$. By combining Eq. 5 and Eq. 8, then maximization of $I(Z;V)$ can be expressed as:

$$\begin{aligned} \max I(Z;V) &= \max \mathbb{E}_{p(z,v)}\Big[\log \frac{p(z|v)}{p(v)}\Big] \\ &\propto \max \sum p(z|v) \log p(v|z) \\ &= \max(-\sum p(z|v) \log \frac{1}{p(v|z)}) \\ &= \min(\sum p(z|v) \log \frac{1}{p(v|z)}) \\ &= \min(\sum CE(z|v, v|z)) \end{aligned} \quad (9)$$

here CE represents cross entropy. Accordingly, CoT distillation in Eq. 4 is transformed into the problem outlined in the above equation. This problem can be addressed with a learning-based method. To tackle this issue, we have developed a new MI loss that minimizes cross-entropy of representation features of the rationale generation ($p(z|v)$) and representation features of the label prediction ($p(v|z)$), effectively maximizing MI during CoT distillation process.

3.3 Training Loss

The training loss is given by

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{\text{prediction}} + \alpha_2 \mathcal{L}_{\text{generation}} + \alpha_3 \mathcal{L}_{\text{CE}} \quad (10)$$

where α_1, α_2 and α_3 are regularization parameters, all of which are non-negative. $\mathcal{L}_{\text{prediction}}$ represents the loss of the label prediction task, and $\mathcal{L}_{\text{generation}}$ represents the loss of the rationale generation task. Both are general cross-entropy loss as defined in (Hsieh et al., 2023).

According to the last line of Equation 9, we define the our MI loss as

$$\mathcal{L}_{\text{CE}} = l(f(\mathbf{Z}), f(\mathbf{V})) \quad (11)$$

f represents our proposed mutual information (MI) loss module, and l denotes the cross-entropy loss. As shown in Figure 1, the MI loss module consists of softmax and max reduction layers. The softmax function separately calculates the distributions for the outputs of the vocabulary spaces in the label prediction and rationale generation tasks. Subsequently, a max reduction operation is employed to reduce the dimensionality of the predicted outputs from both tasks to a uniform dimension for the loss calculation. Specifically, in the label prediction task, dimensions are reduced from $\mathbb{R}^{m\times d}$ to $\mathbb{R}^{1\times d}$, and in the rationale generation task, from $\mathbb{R}^{n\times d}$ to $\mathbb{R}^{1\times d}$.

4 Experiments

4.1 Experimental Setting

Datasets. We conducted the experiments on four widely-used benchmark datasets across three different NLP tasks: e-SNLI (Camburu et al., 2018) and ANLI (Nie et al., 2020) for natural language inference; CQA (Talmor et al., 2018) for commonsense question answering; and SVAMP (Patel et al., 2021) for arithmetic math word problems. We used rationale generated by PaLM 540B (Chowdhery et al., 2023), which were collected and open-sourced by (Hsieh et al., 2023)[1].

Setup. Based on CoT properties and the comparative experimental study in (Hsieh et al., 2023), our work adopted T5-base (220 million) and T5-small (60 million) to the student models. α_1 and α_2 were set as 0.5 and α_3 is set as 0.1. We trained our models on one A100 GPU with $80G$ memory. For T5 base model, the training time for the full-size four dataset was approximately 14.4 hours. For T5 small model, the training times was approximately 8.6 hours.

Baselines. We compare our work with the state-of-the-art DSS (Hsieh et al., 2023) by running their open-sourced code and include two other baseline reported in their work: (1) Standard Finetune, which involves using the prevailing pretrain-then-finetune paradigm to finetune a model with ground-truth labels through standard label supervision. (Howard and Ruder, 2018). (2) Single-task, which finetunes the model using both of the label and non-CoT rationale as supervision .

Evaluation Settings. Following the DSS work (Hsieh et al., 2023), we adopt the accuracy as the performance metrics across all four datasets. Higher accuracy indicates better results. Besides accuracy, we also adopt Expected Calibration Errors (ECE) and Average Confidence Scores to evaluate calibration of the T5-base model. A lower ECE and higher Average Confidence Scores indicate better performance. We adopt GPT-4 to evaluate Quality of CoT examples and subjective analysis. Please refer to our codes for more details.

4.2 Results

Experiments of T5-base Model. We present our experimental results for the T5-base model in Table 1. In single-task training, the rationale and label are concatenated into a single sequence, which is treated as the target in training models (Hsieh et al., 2023). Our proposed method consistently achieves better performance than standard fine-tuning and single-task methods across all datasets. Compared to DSS, our method outperforms DSS on ANLI, CQA, and SVAMP, and achieves nearly the same accuracy on e-SNLI.

[1]Data and DSS code are from https://github.com/google-research/distilling-step-by-step.

	e-SNLI	ANLI	CQA	SVAMP
Standard FT	88.38	43.58	62.19	62.63
Single-task	88.88	43.50	61.37	63.00
DSS	**89.51**	49.58	63.29	65.50
Ours	89.50	**51.20**	**63.88**	**68.00**

Table 1: CoT distillation results on T5-base model. The results of Standard Fine-tune (FT), single-task and DSS methods are from (Hsieh et al., 2023).

	e-SNLI	ANLI	CQA	SVAMP
Standard FT	82.90	42.00	43.16	45.00
DSS	**83.43**	42.90	43.24	48.00
Ours	83.23	**43.70**	**43.90**	**52.50**

Table 2: CoT distillation results on T5-small model.

Model	e-SNLI	ANLI
DSS	82.65	42.80
Ours	82.81	45.50

Table 3: Results on two dataset on T5-base model with LLM generated labels.

Experiments of T5-small Model. The experimental results for the T5-small model are shown in Table 2. The patterns of the results are similar to those of T5-base. Our proposed method consistently achieves better performance than standard finetuning across all dataset. Compared to DSS, our method outperforms DSS on ANLI, CQA and SVAMP, and is just 0.2% less accuracy on e-SNLI.

Distillation with LLM Labels. We conducted an experiment on e-SNLI and ANLI datasets with T5-base model to evaluate the effect of label quality. We distilled the student models using labels generated by 540B PaLM instead of the ground truth. The results are shown in Table 3. Comparing Table 1 and Table 3, we observe the label quality affects the distillation results in both methods. Even With the noisy LLM labels, our model still outperforms DSS on both datasets.

Distillation with smaller datasets. To evaluate the performance of our models on smaller datasets, we distilled T5-base and T5-small models on various sizes of four datasets and compared to DSS method. The results are shown in Figure. 2 and 3 respectively.

	e-SNLI	ANLI	CQA	SVAMP
Mean	89.34	**51.40**	63.88	66.50
Max	**89.50**	51.20	**63.88**	**68.00**

Table 4: Results of Mean Reduction Vs Maximum Reduction on T5-based model.

4.3 Ablation Study

Effectiveness of Difference Dimension Reduction Method In our proposed MI loss module, we utilize maximum reduction to align the dimensions of different features. Additionally, mean reduction serves as an alternative method for dimension reduction, based on the hypothesis that important features can represent better than average features. In Table 4, we present the results of two different layer of MI module. The results indicate the superiority of the MI module with maximum reduction.

Comparison with KL Divergence KL divergence loss has been extensively utilized in KD tasks,serving as a metric for assessing the similarity between two data distributions (Hinton et al., 2015; Zhang et al., 2022c; Gou et al., 2021; Husain et al., 2024). While KL divergence is widely applied in various KD scenarios, modeling DSS using IB framework has proven to be more accurate than using similarity measures, as discussed in Section 3. To validate our hypothesis, we conducted experiments on T5-base model across all four datasets. As shown in Table 5, our proposed method consistently outperforms the KL divergence approach, demonstrating superior performance.

	e-SNLI	ANLI	CQA	SVAMP
KL Divergence	89.42	42.00	62.49	67.00
Ours	**89.50**	**51.2**	**63.88**	**68.00**

Table 5: Results of KD loss VS our proposed cross entropy loss, on T5-base model.

5 Discussion

5.1 Analysis on T5 Calibration

Calibration measures the alignment between a model's predicted accuracy and its confidence levels. Lee et al. (2022) introduced an innovative perspective on model distillation, positioning the teacher model not only as a source of knowledge but also as a tool for identifying mis-calibration during the training of the student model. This ability to maintain calibration and make reliable predictions is crucial for downstream applications and has been the focus of prior studies (Chen et al., 2023; Lee et al., 2022; Jiang et al., 2021). Here, we apply the Expected Calibration Errors (ECE) and Average Confidence Scores to reflect the alignment between the model's predicted probabilities and the actual outcomes, thereby gauging the reliability and certainty of its predictions. Despite the potential limitations inherent in these metrics, we still employ ECE in our experiments due to its simplicity and popularity, as in previous work on investigating the calibration quality of T5 (Chen et al., 2023; Lee et al., 2022).

We employ a 10-bin-based ECE metric and a softmax-based approach to compute average confidence scores from the test outputs across all four datasets. Given that e-SNLI and ANLI essentially represent the same task, we conduct an out-of-domain experiment by testing the model checkpoint trained on one dataset with the test set of the other. This analysis gives us insights into how well our model generalizes across similar tasks and the robustness of its predictions in out-of-domain scenarios and to assess the calibration quality of the model more comprehensively.

Table 6 presents the results of the distilled model calibration evaluation. Overall, both models report lower ECE and confidence scores on SVAMP and e-SNLI, indicating that these two tasks are more challenging and models are less certain about their prediction. Lower ECE values from our MI-based distillation approach are presented for e-SNLI and ANLI, and their respective out-of-domain tests. Notably, our method achieves an ECE of 4.35 in e-SNLI, significantly lower than DSS's 8.54. However, in SVAMP and CQA, our method records higher ECE, indicating potential areas for improvement in these domains. The trade-off in calibration accuracy in specific tasks like SVAMP and CQA compared to DSS suggests future directions for refining our approach.

Regarding average confidence scores (Conf.), our method generally maintains competitive confidence levels, with notable improvements in e-SNLI and ANLI. In e-SNLI, the confidence is lower (30.06) compared to DSS (34.33), which, combined with a lower ECE, suggests a more realistic confidence estimation. Conversely, in the out-of-domain scenarios for e-SNLI and ANLI, our method shows marginally higher confidence scores

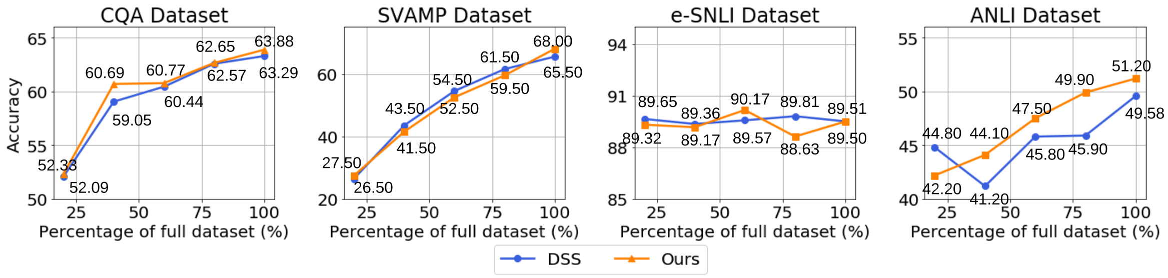

Figure 2: Comparison with DSS with varying sizes of training datasets on T5-base model.

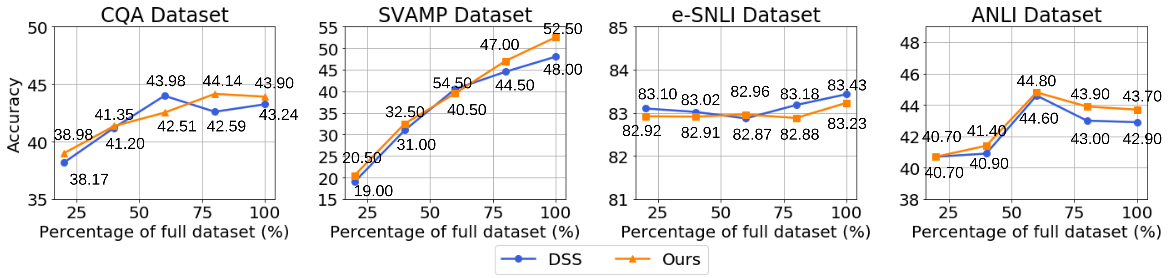

Figure 3: Comparison with DSS with varying sizes of training datasets on T5-small model.

than DSS, which, coupled with the lower ECE, indicates robustness in out-of-domain generalization.

5.2 Analysis on CoT Output

5.2.1 Quality of CoT Examples by GPT-4 Evaluation

We evaluate the quality of CoT examples using GPT-4, as it achieves the state-of-the-art human alignment performance and is used for text generation evaluation in previous work (Liu et al., 2023; Hsu et al., 2023; Wang et al., 2023a). Inspired by (Wang et al., 2023a), we ask GPT-4 to evaluate the quality of the provided CoT examples based on their coherency and relevancy to the input questions and answers. We randomly sample 50 CoT examples from the four datasets and ask GPT-4 to score based on a scale from 1 to 5, where 1 indicates completely incoherent and irrelevant responses, and 5 represents highly coherent, relevant, and helpful responses. For each sample, we run the same sample for four times to obtain self-consistency to measure the reliability of the responses. Table 7 presents the prompt we use for GPT-4 evaluation, average scores and standard deviation on the scores obtained over the four datasets. We report the scores on both provided CoT ("gold") rationales and distilled model predicted rationales.

The effectiveness of our MI-based distillation method is closely linked to the quality of CoT reasoning in the training data. When the CoT quality is high, as in SVAMP, a strong correlation is observed between the model's label prediction accuracy and the quality of its generated CoT. However, this correlation weakens significantly when the CoT quality is low (e-SNLI), suggesting that the model struggles to align label prediction with coherent rationale generation under poor training conditions. Interestingly, with average-quality CoT data (ANLI), the performance gap between our MI-based distillation and DSS is minimal, suggesting that the effectiveness of our approach is particularly reliant on the presence of high-quality reasoning in the training data.

5.2.2 Case Studies on the Output Rationale

We performed case studies on SVAMP and e-SNLI as illustrated in Figure 4 and 5. In the SVAMP example, the question asks the difference in the number of kids Julia played with from Monday to Tuesday, with specific numbers provided for Monday, Tuesday, and Wednesday. DSS generates an incorrect explanation, which contradicts the given question, resulting in to a wrong answer. Conversely, our method correctly identifies the comparison needed between the number of kids Julia played with on Monday and Tuesday, leading to the correct answer. Notably, our generated CoT reason-

Model	SVAMP		CQA		e-SNLI		ANLI		e-SNLI (Out)		ANLI (Out)	
	ECE	Conf.	ECE	Conf.	ECE	Conf.	ECE	Conf.	ECE	Conf.	ECE	Conf.
DSS	11.81	32.56	11.75	42.79	8.54	34.33	11.12	42.72	9.81	38.01	12.78	41.69
Ours	18.92	36.81	13.65	41.17	4.35	30.06	6.94	35.90	6.61	38.08	12.27	42.35

Table 6: Comparisons of our model and DSS on the expected calibration errors (ECE) and average confidence scores (Conf.).

Prompt for GPT 4 Evaluation				
Given an input pair of a question and an answer of a *taskname* task, how good is the given Chain-of-thought example? From 1-5, where 1 is completely incoherent and irrelevant, 2 is somewhat incoherent and irrelevant, 3 is coherent, relevant but not helpful, 4 is somewhat helpful, and 5 is helpful and it explains the answer well.				
Average Scores and Standard Deviation				
Model	**SVAMP**	**CQA**	**e-SNLI**	**ANLI**
Gold	$4.63_{\pm 1.05}$	$3.95_{\pm 1.16}$	$2.42_{\pm 1.23}$	$3.82_{\pm 1.26}$
++	$4.43_{\pm 1.18}$	$4.11_{\pm 1.40}$	$3.49_{\pm 1.35}$	$4.01_{\pm 1.10}$
DSS	$2.50_{\pm 1.42}$	$3.60_{\pm 1.61}$	$3.24_{\pm 1.27}$	$3.48_{\pm 1.40}$
++	$2.53_{\pm 1.46}$	$3.64_{\pm 1.62}$	$3.18_{\pm 1.21}$	$3.44_{\pm 1.30}$
Ours	$2.30_{\pm 1.54}$	$3.70_{\pm 1.45}$	$3.03_{\pm 1.47}$	$3.42_{\pm 1.37}$
++	$2.72_{\pm 1.45}$	$3.63_{\pm 1.60}$	$3.17_{\pm 1.17}$	$3.34_{\pm 1.21}$

Table 7: Prompt used and results of 50 randomly sampled CoT examples from the four datasets evaluated by GPT-4. We use ++ to denote the setting with *self-consistency* evaluation.

Model	SVAMP	CQA	e-SNLI	ANLI
DSS	0.12	0.66	0.05	0.26
	$p > 0.05$	$p < 0.05$	$p > 0.05$	$p > 0.05$
Ours	0.42	0.53	0.03	0.26
	$p < 0.05$	$p < 0.05$	$p > 0.05$	$p > 0.05$

Table 8: Pearson correlation between CoT quality and accuracy of label prediction on the 50 random samples on the test set. We highlight the correlation with statistical significance ($p < 0.05$).

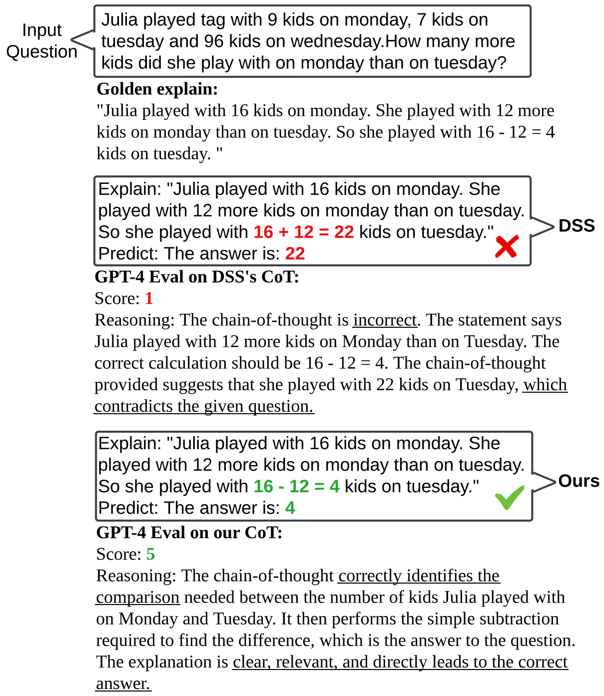

Figure 4: A case study of the output rationale on SVAMP dataset.

ing is identical to the golden one, demonstrating that by precisely grasping the rationale, our approach effectively resolves the math problem. We also show the evaluation results (score and reasoning) from GPT-4, where our method gains a top score of 5 and DSS gains only a mere score of 1. This example showcases that the high-quality CoT generated by our method enhances problem-solving capabilities in math tasks like SVAMP.

Another example (Figure 5) is from e-SNLI, where the task is to identify whether the hypothesis is entailment, contradiction, or neutral, based on the given premise and hypothesis. Although both our method and DSS generate the correct label output, it is worth noting that, the CoT of our method points out the relationship between the premise and the hypothesis, while DSS only restates the hypothesis without providing any extra explanation or connecting the hypothesis to the premise. Our generated rationale also gains a higher score than DSS. A higher-quality rationale tends to facilitate more accurate label prediction, thereby enhancing overall task performance.

6 Conclusion

In this paper, we re-investigate the DSS framework from an information-theoretic perspective. We model it using Information Bottleneck and propose to strengthen it by maximizing the mutual information between rationale generation and label prediction tasks. The proposed learning-based method can automatically optimize the CoT distillation and bolster the reasoning ability of the distilled smaller models. Both our qualitative and quantitative analysis demonstrate the rationale behind

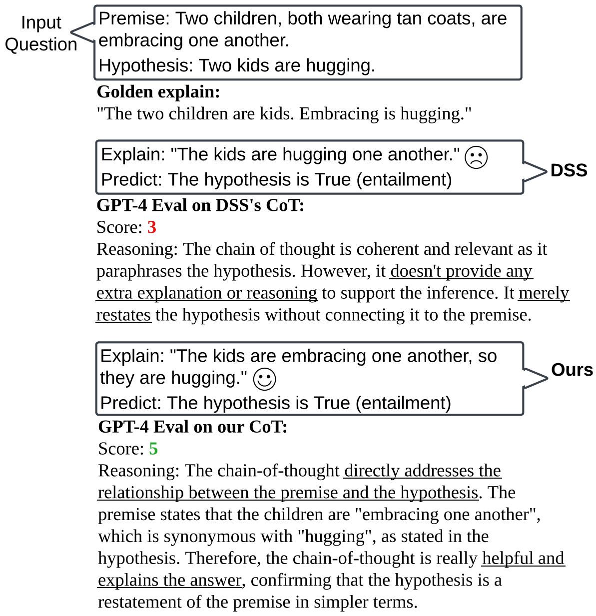

Figure 5: A case study of the output rationale on e-SNLI dataset.

our method and shed light on aspects of language model distillation and CoT applications.

7 Limitation

Our comparative analysis primarily focuses on the Distilling Step-by-Step (DSS) framework, which serves as our main benchmark. This concentrated comparison, while valuable for a deep understanding of DSS's nuances and our advancements over it, constitutes a limitation of our work. Specifically, our analysis does not extend to a broader range of knowledge distillation methods currently employed in the field, focusing exclusively on T5 and not including other LLMs like Mistral and Llama2, Llama3 model family. This focus may overlook the potential insights and contrasts that could emerge from evaluating our approach against a wider array of distillation techniques. Future research could benefit from a more expansive comparative study, incorporating diverse methodologies to fully contextualize our findings within the broader landscape of knowledge distillation practices. This broader comparison would not only validate the efficacy of our method in various settings but also illuminate areas for further refinement and innovation.

However, it is important to note that our contribution lies in providing an in-depth analysis from both theoretical and practical viewpoints to enhance the CoT distillation process. Our work delves into the intricacies of utilizing mutual information to improve distillation outcomes, offering significant advancements in understanding and applying CoT distillation techniques.

8 Ethical Issues

In this paper, we carefully considered the ethical implications in line with the ACL code of ethics. We evaluated the potential dual-use concerns, ensuring our research serves to benefit society and does not cause inadvertent harm. Our methodology and applications were thoroughly assessed for fairness, non-discrimination, and privacy, particularly in the context of data handling and model outputs. We also ensured our study did not expose any negative impact on individuals and groups. Moreover, we did not engage in academic dishonesty and adhered to high-quality processes and product standards in our professional work. We include this detailed discussion of these ethical considerations, affirming our commitment to responsible and beneficial computational linguistics research.

Acknowledgements

We thank reviewers for providing helpful feedback.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. In *International Conference on Learning Representations*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2023. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*.

Philip Bachman, R Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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

Hanjie Chen and Yangfeng Ji. 2020. Learning variational word masks to improve the interpretability of neural text classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251.

Pengguang Chen, Shu Liu, Hengshuang Zhao, and Jiaya Jia. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5008–5017.

Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. 2023. A close look into the calibration of pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1343–1367, Toronto, Canada. Association for Computational Linguistics.

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

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

Ian Connick Covert, Wei Qiu, Mingyu Lu, Na Yoon Kim, Nathan J White, and Su-In Lee. 2023. Learning to maximize mutual information for dynamic feature selection. In *International Conference on Machine Learning*, pages 6424–6447. PMLR.

Michael Ziyang Diao, Krishna Balasubramanian, Sinho Chewi, and Adil Salim. 2023. Forward-backward gaussian variational inference via jko in the bures-wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR.

Yingjun Du, Jun Xu, Huan Xiong, Qiang Qiu, Xiantong Zhen, Cees GM Snoek, and Ling Shao. 2020. Learning to learn with variational information bottleneck for domain generalization. In *ECCV 2020*, pages 200–216. Springer.

Kelly G Garner and Paul E Dux. 2023. Knowledge generalization and the costs of multitasking. *Nature Reviews Neuroscience*, 24(2):98–112.

Dongyoung Go, Tomasz Korbak, Germán Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. 2023. Aligning foundation models for language with preferences through f-divergence minimization. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129:1789–1819.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Ting-Yao Hsu, Chieh-Yang Huang, Ryan Rossi, Sungchul Kim, C Giles, and Ting-Hao Huang. 2023. Gpt-4 as an effective zero-shot evaluator for scientific figure captions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5464–5474.

Hisham Husain, Vu Nguyen, and Anton van den Hengel. 2024. Distributionally robust bayesian optimization with $\varphi-$divergences. *Advances in Neural Information Processing Systems*, 36.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. Tinybert: Distilling bert for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Dongkyu Lee, Zhiliang Tian, Yingxiu Zhao, Ka Chun Cheung, and Nevin Zhang. 2022. Hard gate knowledge distillation-leverage calibration for robust and reliable language model. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9793–9803.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*.

Wei-Hong Li and Hakan Bilen. 2020. Knowledge distillation for multi-task learning. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 163–176. Springer.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. 2019. Pareto multi-task learning. *Advances in neural information processing systems*, 32.

Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang. 2009. Feature selection with dynamic mutual information. *Pattern Recognition*, 42(7):1330–1339.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. 2024. Diff-instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *Advances in Neural Information Processing Systems*, 36.

Yuhan Ma, Haiqi Jiang, and Chenyou Fan. 2023. Sci-cot: Leveraging large language models for enhanced knowledge distillation in small models for scientific qa. *arXiv preprint arXiv:2308.04679*.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. 2023. Teaching small language models to reason. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1773–1781, Toronto, Canada.

David McAllester and Karl Stratos. 2020. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR.

David Mueller, Nicholas Andrews, and Mark Dredze. 2022. Do text-to-text multi-task learners suffer from task conflict? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2843–2858.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Liam Paninski. 2003. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Mary Phuong and Christoph Lampert. 2019. Towards understanding knowledge distillation. In *International conference on machine learning*, pages 5142–5151. PMLR.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. 2019. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Fynn Schröder and Chris Biemann. 2020. Estimating the influence of auxiliary tasks for multi-task learning of sequence tagging tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2971–2985.

Noam Slonim. 2002. *The information bottleneck: Theory and applications*. Ph.D. thesis, Hebrew University of Jerusalem Jerusalem, Israel.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Xudong Tian, Zhizhong Zhang, Shaohui Lin, Yanyun Qu, Yuan Xie, and Lizhuang Ma. 2021. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1522–1531.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *International Conference on Learning Representations*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. IEEE.

Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. 2019. On mutual information maximization for representation learning. In *International Conference on Learning Representations*.

Zhibin Wan, Changqing Zhang, Pengfei Zhu, and Qinghua Hu. 2021. Multi-view information-bottleneck representation learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10085–10092.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Lin Wang and Kuk-Jin Yoon. 2021. Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3048–3068.

Peifeng Wang, Zhengyang Wang, Zheng Li, Yifan Gao, Bing Yin, and Xiang Ren. 2023b. SCOTT: Self-consistent chain-of-thought distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5546–5558, Toronto, Canada. Association for Computational Linguistics.

Qi Wang, Claire Boudreau, Qixing Luo, Pang-Ning Tan, and Jiayu Zhou. 2019. Deep multi-view information bottleneck. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 37–45. SIAM.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Joseph Worsham and Jugal Kalita. 2020. Multi-task learning for natural language processing in the 2020s: where are we going? *Pattern Recognition Letters*, 136:120–126.

Yangyang Xu, Yibo Yang, and Lefei Zhang. 2023. Multi-task learning with knowledge distillation for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21550–21559.

Chenxiao Yang, Junwei Pan, Xiaofeng Gao, Tingyu Jiang, Dapeng Liu, and Guihai Chen. 2022. Cross-task knowledge distillation in multi-task recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 4318–4326.

Cenyuan Zhang, Xiang Zhou, Yixin Wan, Xiaoqing Zheng, Kai-Wei Chang, and Cho-Jui Hsieh. 2022a. Improving the adversarial robustness of nlp models by information bottleneck. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3588–3598.

Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. 2021. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4388–4403.

Linfeng Zhang, Xin Chen, Xiaobing Tu, Pengfei Wan, Ning Xu, and Kaisheng Ma. 2022b. Wavelet knowledge distillation: Towards efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12464–12474.

Linfeng Zhang, Xin Chen, Junbo Zhang, Runpei Dong, and Kaisheng Ma. 2022c. Contrastive deep supervision. In *European Conference on Computer Vision*, pages 1–19. Springer.

Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, and Kede Ma. 2023a. Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14071–14081.

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

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