Leave the Bias in Bias: Mitigating the Label Noise Effects in Continual Visual Instruction Fine-Tuning

Anonymous ACL submission

Abstract

In recent years, multimodal large language models (MLLMs) with vision processing capability have shown substantial advancements, excelling particularly in interpreting general images. Their application in domain-specific tasks, like those in the medical fields, is further enhanced through continuous visual instruction fine-tuning (CVIF). Despite these advancements, a significant challenge arises from label noise encountered during the collection of domain-specific data. Our studies reveal that this label noise can adversely affect the learning of vision projection embeddings and contribute to inaccuracies in LLMs' fine-tuning, often leading to hallucinations. In this paper, we introduce a novel framework designed to minimize the impact of label noise. Our approach focuses on stabilizing the learning of vision embeddings and reducing the effect of label noise through the inherent semantic understanding of uncertainty in LLMs. Extensive experiments demonstrate that our framework maintains robust performance in general visual question-answer (VQA) tasks while showing significant effectiveness in medical VQA tasks. To the best of our knowledge, this is the first study to specifically address and analyze the impact of label noise in CVIF.

1 Introduction

Recent advancements in large language models (LLMs) have significantly advanced artificial general intelligence (AGI) (Touvron et al., 2023; Floridi and Chiriatti, 2020; Chiang et al., 2023). Enhancing LLMs' ability to process multimodal real-world data, particularly integrating visual data, is key to developing universal AGI interfaces that facilitate human interaction (Radford et al., 2021). Studies have focused on using vision-instructed tuning to align visual inputs with semantic representations in LLMs, enabling them to process real-world visual signals (Li et al., 2023b; Zhu et al., 2023; Liu et al., 2023). This alignment enhances the capabilities of AGI assistants, allowing users to interact with and manipulate visual inputs using natural language commands.

To improve the performance of multimodal large language models (MLLMs) in specific domains, it's crucial to apply continuous multimodal instruction fine-tuning using tailored datasets (Zhang et al., 2023a; Yan et al., 2021; Wu et al., 2023). This technique is also vital for visual-based LLMs, where continual vision instruction fine-tuning (CVIF) leverages domain-specific images and guidance (Li et al., 2023a; Zhang et al., 2023b). However, building these datasets often encounters the challenge of label noise—incorrect or inaccurate labels stemming from data annotation inconsistencies, automated processing errors, or subjective human judgment, especially in complex areas like medicine where expert interpretations vary (Han et al., 2018; Liu et al., 2020; Xia et al., 2020; Liu et al., 2021b).

Label noise in instructional data induces hallucinations in LLMs post fine-tuning, notably in text-based LLMs (Qi et al., 2023; Dong et al., 2023) but is under-researched in vision-based models, particularly in general visual question-answer (VQA) tasks. Section B of our study reveals that label noise not only triggers hallucinations in vision-based LLMs but also adversely affects the projection layer, crucial for visual interpretation (Li et al., 2023b; Zhu et al., 2023; Liu et al., 2023). This dual impact significantly biases inference. Prior studies mainly consider label noise in classification tasks using label transition matrices (Zhang and Sabuncu, 2018; Wang et al., 2017; Chen and Gupta, 2015; Yong et al., 2022), which are insufficient for the nuanced demands of VQA tasks.

We propose a novel framework targeting VQA tasks that mitigates label noise by focusing on both the **projection layer** and **LLMs**. We first employ Polyak averaging techniques (Polyak, 1964) to reduce the overfitting of bias in projection layers.

Concurrently, we leverage the inherent bias understanding of certain phrases within LLMs (Zhou et al., 2023), which learned from extensive text corpus learning, to learn the bias instruction data and infer uncertainly to mitigate the label noise influence. Hence, we term our framework "**Lea**ve the **B**ias i**n** **B**ias" (LEABNB). We conduct experiments in general VQA tasks using the Llava model (Liu et al., 2023) and domain-specific medical VQA tasks employing MedVInT_TD (Zhang et al., 2023b). Experimental results indicate that LEABNB demonstrates significant robustness to label noise and achieves performance comparable to the standard fine-tuning process when applied to clean datasets. This makes LEABNB well-suited for general CVIF scenarios. Our contributions are concluded as follows:

- We first investigate the label noise effect of vision-based LLMs in CVIF.
- Building upon these findings, we introduce LEABNB, a novel framework designed to effectively mitigate the effects of label noise in CVIF.
- Our methodology validates and leverages the LLMs' inherent bias understanding of certain phrases in label noise reduction.
- To the best of our knowledge, LEABNB represents the first framework for mitigating the effect of label noise in CVIF.

2 Method

2.1 Preliminary

Vision-based LLMs integrate pre-trained textual LLMs (Chiang et al., 2023; Touvron et al., 2023) with visual models (Radford et al., 2021), which are initially trained on distinct datasets for text and images. For applications like VQA, it is crucial to align visual information with the text-based knowledge of LLMs. A standard pre-trained LLM, p_θ, undergoes fine-tuning with an instruction dataset $\mathcal{D}_{IF}$, comprising instruction-response pairs $(\mathbf{x}, \mathbf{y})$. This process aims to maximize the log-likelihood of generating correct responses, formulated as:

$$\mathbb{E}_{\mathcal{D}_{IF}} \log p_\theta(\mathbf{y}) = \mathbb{E}_{\mathcal{D}_{IF}} \log \prod_{i=1}^{k} p_\theta(y_i|\mathbf{x}), \quad (1)$$

Additionally, techniques like RLHF (Ouyang et al., 2022) are utilized to enhance alignment with human instructions and promote the LLMs' helpfulness, harmlessness, and honesty.

To enable LLMs to process visual information, we employ a pre-trained visual model, specifically CLIP (Radford et al., 2021), to generate visual embeddings $\mathbf{z}_e = g(\mathbf{z})$. These embeddings are integrated into LLMs using a projection encoder h_γ and a vision instruction fine-tuning (VIF) dataset $\mathcal{D}_V$, which consists of tuples $(\mathbf{z}, \mathbf{x}, \mathbf{y})$. The vision-augmented LLM, denoted as $\pi_{\theta,\gamma}$, merges parameters θ from the LLM and γ from the projection layer. Optimization is achieved by maximizing the model's log-likelihood:

$$\mathbb{E}_{\mathcal{D}_V} \log \pi_{\theta,\gamma}(\mathbf{y}) = \mathbb{E}_{\mathcal{D}_V} \log \prod_{i=1}^{k} p_\theta(y_i|h_\gamma(\mathbf{z}_e), \mathbf{x}), \quad (2)$$

where the expectation is calculated over the VIF dataset, considering output tokens y_i, conditioned on projected embeddings and input tokens. For domain-specific applications like medical VQA, we adapt the LLM using a specialized dataset $\mathcal{D}_C$, following Equation (2).

However, a critical issue in CVIF is the introduction of label noise during the learning process. Label noise, which stems from human or machine errors in dataset labels (Algan and Ulusoy, 2021), can significantly undermine the learning process. For instance, a benign skin lesion in a medical dataset might be mislabeled as malignant due to diagnostic inaccuracies (Liu et al., 2021b). Such errors can cause inconsistencies across datasets and conflict with the intrinsic knowledge of LLMs. Consequently, when $\pi_{\theta,\gamma}$ is fine-tuned with these datasets, it may destabilize the learning process and lead to misinterpretations of medical images, a critical concern in healthcare.

2.2 Leave the Bias in Bias

By analyzing the impact of label noise in Appendix B, the label noise can influence both θ and γ, and induce compounded performance drop. Here, we present the LEABNB framework to mitigate label noise influence in CVIF, which is shown in Figure 1. Based on the case study results shown in the Table 2, we can initially froze the projection layer to mitigate overfitting from noisy labels, which proved somewhat effective. However, this approach limited the model's ability to generalize to new tasks. Consequently, we adopted Polyak Averaging for the projection layer γ, employing a decay factor λ to modulate updates, as illustrated in Equation

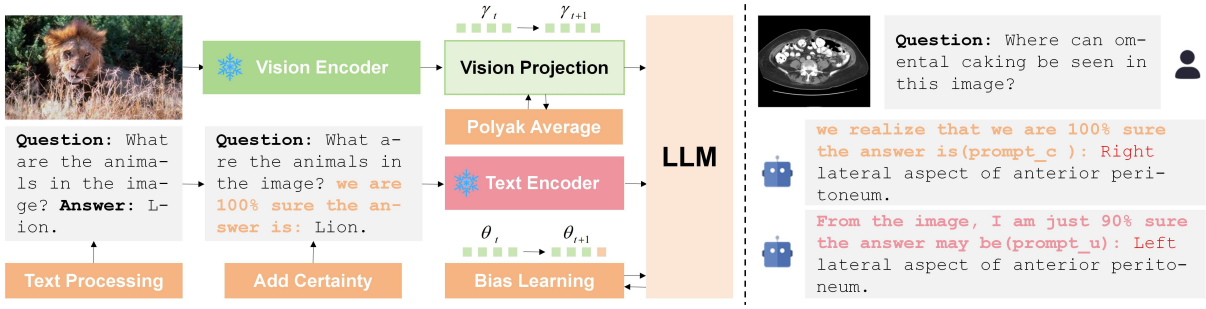

Figure 1: The CVIF phase for vision-based LLMs, depicted on the left, incorporates a specific preprocessing protocol. We use a graphical template to format "Questions" and "Answers" for VQA tasks and initiate with a biased, leading prompt to evoke an overly confident response from the model. The Polyak averaging strategy manages the weight updates in the projection layer. The inference phase, shown on the right, demonstrates that using $prompt_c$ (e.g., "100% confidence") leads to biased and often incorrect responses. Conversely, employing $prompt_u$ (e.g., "90% confidence") helps the model shed its biases and produce more accurate answers.

(3). At each iteration $t+1$, the parameters are updated by combining them, scaled by $\lambda = 0.05$, with the parameters from the previous time step t. This method not only preserves the model's generalization across new tasks but also counters the detrimental effects of noisy data.

$$\bar{\gamma}_{t+1} = \lambda\bar{\gamma}_{t+1} + (1-\lambda)\gamma_t, \tag{3}$$

where λ tuning the update extent and γ_{t+1} updated by Equation (2).

In the field of LLMs, which are generative rather than typical classification-based, traditional methods (Radford et al., 2019) for handling label noise are ineffective. This is because LLMs depend on semantic comprehension rather than purely probabilistic learning on labeled data. Our proposed framework, LEABNB, seeks to mitigate label noise in LLMs by exploiting their semantic capabilities. Several studies (Laranjo et al., 2018; Kadavath et al., 2022; Zhou et al., 2023) have shown LLMs may develop biased reasoning from phrases with overconfident meanings, a bias rooted in human language's unique features that can be observed in our daily lives. For example, the human may utilize phrase "I'm 100% certain..." with false statements due to overconfidence. Also, such phrase often implies negation in LLMs' pre-training corpora (e.g., "I'm not 100% sure"), which can also bias LLMs reasoning.

Inspired by (Liu et al., 2022) that utilizes overparameterization to handle label noise by assigning a specific output parameter to each data point and then mitigating corrupted label noise through inference without these parameters, we introduce a new strategy within our LEABNB framework, termed **Bias-learning**. This method employs the prompt $prompt_c$, characterized by overconfidence and certainty, to update the model parameter θ. In contrast, during inference, we employ prompts $prompt_u$ that are uncertain and conservative. The name of our method, "Leave the bias in bias," reflects its purpose. In QA tasks, using overly confident phrases during model fine-tuning can introduce biases. These biases originate from the corpus knowledge acquired in the pre-training phase, often associated with expressions of negation and incorrect answers. By introducing uncertainty during the inference stage, we can modify the semantic environment, enabling LLMs to confine the biases they have learned within specific, overconfident prompt environments, and encourage them to re-reason their answers. As fine-tuning activates existing capabilities without adding new knowledge, correct labeling remains unaffected, and the issue of hallucinations due to overfitting on incorrect labels is addressed. Consequently, the model parameters at time step t are updated according to the following sequential rules:

$$\begin{aligned}
\theta_{t+1} &= \theta_t + \alpha\nabla_\theta \mathbb{E}_{\mathcal{D}_V} \log p_\theta(\mathbf{y}|h_{\bar{\gamma}_t}(\mathbf{z}_e), \mathbf{x}_c, \\
\gamma_{t+1} &= \bar{\gamma}_t + \alpha\nabla_\gamma \mathbb{E}_{\mathcal{D}_V} \log p_\theta(\mathbf{y}|h_{\bar{\gamma}_t}(\mathbf{z}_e), \mathbf{x}_c, \\
\bar{\gamma}_{t+1} &= \lambda\bar{\gamma}_t + (1-\lambda)\gamma_{t+1}, \mathbf{x}_c = prompt_c(\mathbf{x}))
\end{aligned} \tag{4}$$

where α represents the learning rate. After the training, we perform the VQA task by sampling model answers $\mathbf{y}'$ with uncertain prompts: $\mathbf{y}' \sim p_\theta(\cdot|h_\gamma(\mathbf{z}_e), prompt_u(\mathbf{x}))$.

3 Experiment

To evaluate our method's efficacy, we experimented on four recognized datasets and designed three experimental settings: standard SFT training, MW-

Table 1: The performance of LEABNB on four datasets

Task	Method	0%	10%	20%	40%
Slake	SFT	74.45 ± 0.62	70.53 ± 0.54	65.96 ± 1.02	45.42 ± 0.83
	MW-Net	-	71.32 ± 0.33	66.82 ± 0.89	47.46 ± 0.23
	ours	-	$\mathbf{71.82 \pm 0.51}$	$\mathbf{68.53 \pm 0.63}$	$\mathbf{51.67 \pm 0.92}$
VQA-RAD	SFT	52.02 ± 0.53	48.84 ± 0.81	46.46 ± 0.41	37.25 ± 0.73
	MW-Net	-	49.82 ± 0.53	47.62 ± 1.23	38.86 ± 0.63
	ours	-	$\mathbf{50.17 \pm 0.43}$	$\mathbf{48.48 \pm 0.68}$	$\mathbf{41.69 \pm 1.22}$
GQA	SFT	64.51 ± 0.62	60.84 ± 0.58	54.63 ± 1.33	51.85 ± 0.91
	MW-Net	-	61.72 ± 0.63	55.58 ± 0.66	54.72 ± 0.86
	ours	-	$\mathbf{63.22 \pm 0.73}$	$\mathbf{60.27 \pm 1.29}$	$\mathbf{57.14 \pm 0.97}$
OKVQA	SFT	44.82 ± 1.22	42.76 ± 0.98	39.21 ± 0.75	34.74 ± 0.86
	MW-Net	-	$\mathbf{43.42 \pm 0.63}$	39.74 ± 0.56	34.24 ± 0.94
	ours	-	43.35 ± 0.54	$\mathbf{42.56 \pm 1.29}$	$\mathbf{37.12 \pm 0.49}$

Net, and our approach. MW-Net excels in fine-tuning MLLMs through a meta-network that dynamically adjusts loss function weights to alleviate label noise impacts and prevent overfitting (Huang et al., 2023; Friend et al., 1993; Zhang and Sabuncu, 2018). Other approaches handle label noise by estimating a transition matrix, suitable only for classification with fixed classes (Yao et al., 2020; Yang et al., 2022; Bae et al., 2024). In contrast, MW-Net, a model-agnostic method, recalculates loss values across various models and tasks, serving as a versatile baseline in our experiments. We refer the readers to Appendix C.1 for more information about the dataset preparation and construction for our experiment.

As shown in Table 1, we introduced different levels of label noise into these datasets, specifically including three noise levels of 10%, 20%, and 40%. As the noise ratio increases, the memory effect of deep learning models causes them to fit more incorrect knowledge, resulting in a significant decrease in test accuracy. However, our method exhibits excellent robustness in handling high proportions of noise. As noise levels increase, our method allows the LLM to learn more biased information, which significantly enhances its performance. We refer the readers to Appendix C.2 for the information about evaluation metrics and model hyperparameters for the experiment.

The experimental results demonstrate that LEABNB surpasses the baseline in almost all four datasets. By observing the inference log, the results likely arises from how language model responses under LEABNB, consisting of multiple tokens, are minimally affected by noisy labels that only alter a few key tokens crucial for semantic meaning. The answer of MW-Net, however, struggles with these fine semantic distinctions between clean and noisy labels, failing to adjust weights adequately to prevent overfitting due to noisy data. Specifically, in the specialized medical VQA datasets, our method performed better in the SLAKE dataset than in the VQA-RAD dataset, likely due to differences in handling open-ended questions. VQA-RAD's lengthier and more complex answers add to the prediction challenges. Conversely, in the general VQA datasets, our approach showed superior performance on the GQA dataset compared to the OKVQA dataset because GQA questions depend solely on the image and its contents, whereas OKVQA requires integrating extensive external knowledge, complicating the model's ability to capture accurate answers.

To further evaluate our proposed LEABNB, we ablate the effectiveness of Polyak Averaging and Bias-learning to improve its robustness against noisy data. The ablation results are shown in the Appendix D.

4 Conclusion

In this study, we observe that label noise detrimentally affects both the visual and language modules of vision-based LLMs during the CVIF process, reducing model performance. We introduce LEABNB, a novel framework designed to mitigate label noise in vision-based LLMs by using Polyak Averaging for enhanced stability in the projection layer and employing bias learning to utilize LLMs' inherent deterministic semantic understanding. We tested LEABNB on two popular vision-based LLMs across four open-source benchmark datasets, where it demonstrated substantial improvements in general and domain-specific VQA tasks. To the best of our knowledge, LEABNB is the first framework specifically aimed at counteracting label noise during the CVIF process.

5 Limitations

In this study, we encountered several limitations that can be addressed in the future. First, our experiments were exclusively conducted using vision-based LLMs on the Llama (Touvron et al., 2023) and Vicuna (Chiang et al., 2023) with 7B parameter sizes, leaving the impact of our research framework on LLMs with large parameter scales as area for future exploration. Second, we limited our investigation to four datasets, observing varied performances of our framework across different general and domain-specific VQA datasets. This variation highlights the need for further research to assess the generalizability of our framework across a wider range of VQA datasets from diverse domains. Lastly, we discovered that the design of the prompts plays a crucial role in influencing the results of the inference process during learning. Therefore, future studies will focus on developing strategies for designing optimal and stable prompts to improve inferential effectiveness.

References

Malak Abdullah, Alia Madain, and Yaser Jararweh. 2022. Chatgpt: Fundamentals, applications and social impacts. In *2022 Ninth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. Ieee.

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Görkem Algan and Ilkay Ulusoy. 2021. Image classification with deep learning in the presence of noisy labels: A survey. *Knowledge-Based Systems*, 215:106771.

Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren. 2019. Robust bi-tempered logistic loss based on bregman divergences. *Advances in Neural Information Processing Systems*, 32.

Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. 2021. Noise estimation using density estimation for self-supervised multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6644–6652.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR.

HeeSun Bae, Seungjae Shin, Byeonghu Na, and Il-Chul Moon. 2024. Dirichlet-based per-sample weighting by transition matrix for noisy label learning. *arXiv preprint arXiv:2403.02690*.

Xinlei Chen and Abhinav Gupta. 2015. Webly supervised learning of convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1431–1439.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Guanting Dong, Jinxu Zhao, Tingfeng Hui, Daichi Guo, Wenlong Wang, Boqi Feng, Yueyan Qiu, Zhuoma Gongque, Keqing He, Zechen Wang, et al. 2023. Revisit input perturbation problems for llms: A unified robustness evaluation framework for noisy slot filling task. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 682–694. Springer.

Luciano Floridi and Massimo Chiriatti. 2020. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694.

Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on information systems (TOIS)*, 14(3):330–347.

Marilyn Friend, Monica Reising, and Lynne Cook. 1993. Co-teaching: An overview of the past, a glimpse at the present, and considerations for the future. *Preventing School Failure: Alternative Education for Children and Youth*, 37(4):6–10.

Jacob Goldberger and Ehud Ben-Reuven. 2016. Training deep neural-networks using a noise adaptation layer. In *International conference on learning representations*.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Runhui Huang, Yanxin Long, Jianhua Han, Hang Xu, Xiwen Liang, Chunjing Xu, and Xiaodan Liang. 2023. Nlip: Noise-robust language-image pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 926–934.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Liliana Laranjo, Adam G Dunn, Huong Ly Tong, Ahmet Baki Kocaballi, Jessica Chen, Rabia Bashir, Didi Surian, Blanca Gallego, Farah Magrabi, Annie YS Lau, et al. 2018. Conversational agents in healthcare: a systematic review. *Journal of the American Medical Informatics Association*, 25(9):1248–1258.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023c. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021a. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiarun Liu, Ruirui Li, and Chuan Sun. 2021b. Co-correcting: noise-tolerant medical image classification via mutual label correction. *IEEE Transactions on Medical Imaging*, 40(12):3580–3592.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. 2022. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pages 14153–14172. PMLR.

Tongliang Liu and Dacheng Tao. 2015. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461.

Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *International conference on machine learning*, pages 6543–6553. PMLR.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. 2019. Can gradient clipping mitigate label noise? In *International Conference on Learning Representations*.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1944–1952.

Boris T Polyak. 1964. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

Zhenting Qi, Xiaoyu Tan, Chao Qu, Yinghui Xu, and Yuan Qi. 2023. Safer: A robust and efficient framework for fine-tuning bert-based classifier with noisy labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 390–403.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. 2019. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Meiqi Sun, Wilson Yan, Pieter Abbeel, and Igor Mordatch. 2022. Quantifying uncertainty in foundation models via ensembles. In *NeurIPS 2022 Workshop on Robustness in Sequence Modeling*.

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

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Ruxin Wang, Tongliang Liu, and Dacheng Tao. 2017. Multiclass learning with partially corrupted labels. *IEEE transactions on neural networks and learning systems*, 29(6):2568–2580.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 322–330.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. 2020. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. 2021. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. 2022. Estimating instance-dependent bayes-label transition matrix using a deep neural network. In *International Conference on Machine Learning*, pages 25302–25312. PMLR.

Yu Yao, Tongliang Liu, Bo Han, Mingming Gong, Jiankang Deng, Gang Niu, and Masashi Sugiyama. 2020. Dual t: Reducing estimation error for transition matrix in label-noise learning. *Advances in neural information processing systems*, 33:7260–7271.

LIN Yong, Renjie Pi, Weizhong Zhang, Xiaobo Xia, Jiahui Gao, Xiao Zhou, Tongliang Liu, and Bo Han. 2022. A holistic view of label noise transition matrix in deep learning and beyond. In *The Eleventh International Conference on Learning Representations*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhilu Zhang and Mert Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Related work

A.1 Multimodal Large Language Models

LLMs have driven transformative advancements in artificial intelligence and related fields. For instance, ChatGPT (Abdullah et al., 2022), leveraging cutting-edge techniques such as instructional fine-tuning (Li et al., 2023c; Liu et al., 2023; Ouyang et al., 2022) and reinforcement learning from human feedback (RLHF) (Stiennon et al., 2020), has demonstrated exceptional capabilities in language understanding and logical reasoning. Since the introduction of GPT-4 (Achiam et al., 2023), researchers have explored its significant multimodal capabilities. Multimodal learning involves mapping data from different modalities (i.e., text, images, and audio.) to a shared representational space, enabling comparison, association, or joint processing of data from these various sources.CLIP (Radford et al., 2021) employs contrastive learning to map images and text into a shared vector space, thereby improving the semantic alignment between visual and linguistic descriptions. Subsequently, GIT (Wang et al., 2022) refined spatial alignment between images and text, while BLIP2 (Li et al., 2023b) improved the efficiency of pre-trained models, optimizing their performance. Flamingo (Alayrac et al., 2022) leveraged unsupervised pre-training on a large scale of unannotated multimodal data, making it adept at visual tasks with limited annotated resources. MiniGPT-4 (Zhu et al., 2023) improved text generation through a training strategy using self-generated data. LLaVA (Liu et al., 2023) achieved a mapping of visual and textual information to same-dimensional embeddings, with updates in model weights and fine-tuning significantly boosting performance on complex semantic tasks.

A.2 The Impact of Uncertainty and Overconfidence on Language Models

Early research on biases in language models and computer systems (Friedman and Nissenbaum, 1996) revealed that such systems could exhibit bias due to inherent data prejudices. This research underscored the importance of understanding and addressing uncertainties and overconfidence in systems.

With advancements in deep learning in the 21st century, more attention has been given to issues of uncertainty and overconfidence in language models. However, earlier efforts primarily focused on improving the accuracy of model confidence estimates(Sun et al., 2022; Kuhn et al., 2023), precisely measuring uncertainty(Kwiatkowski et al., 2019; Liang et al., 2022), and optimizing calibration performance. These studies adopted a multi-dimensional approach ranging from model ensembles to fine-tuning single-model details, aiming to refine the models' recognition and expression of predictive uncertainty.On the other hand, some studies have examined the impact of certainty in language expressions on model performance. Mielke et al.(Mielke et al., 2022) proposed a solution to reduce model overconfidence through linguistic calibration. Their research prioritized improving the model's certainty in responses to better reflect the accuracy of its answers. Kadavath et al.(Kadavath et al., 2022) experimented with the model's ability to express confidence after determining the correctness of its answers and found the model to be relatively well-calibrated in various scenarios. Their work further demonstrated the model's potential in self-assessing the accuracy of its expressions. Recent work(Zhou et al., 2023) has revealed the high sensitivity of Large Language Models (LLMs) to certainty, uncertainty, or evidential language prompts, indicating that prompts with extreme certainty might compromise model performance, while those containing uncertainty or evidential cues could enhance it. This finding is significant for optimizing the model's input processing mechanisms, showing potential for promoting output accuracy through fine-tuning input prompts.

A.3 Label Noise

Learning from noisy data has been a persistent area of focus for researchers aiming to mitigate the impact of noise in training data, primarily concentrating on classification tasks. Existing studies, such as Song et al.(Song et al., 2022), typically employ robust architectural designs, regularization techniques, loss function adjustments, or sample selection strategies to suppress the adverse effects of noisy labels.

Here, we discuss several popular works. Robust loss functions(Zhang and Sabuncu, 2018; Wang et al., 2019; Amid et al., 2019; Ma et al., 2020) are among the most prevalent methods for addressing label noise, aiming to reduce the loss impact of outliers and thereby alleviate the effects of label noise. Similar concepts are also present in gradient clipping(Menon et al., 2019) and loss reweighting strategies(Liu and Tao, 2015; Wang et al., 2017).

Among them, Meta-weight-net(Shu et al., 2019) is a meta-learning method that aims to improve the robustness of models on noisy labeled data by learning sample weights. It introduces a meta-network to predict the weight of each training sample and alternately optimizes with the main network. However, this method still faces challenges in practical applications, such as the difficulty in designing and optimizing the meta-network, high computational overhead, and sensitivity to the distribution of noisy labels.Another method to handle label noise operates on the assumption that noise labels are generated according to the conditional probability distribution of the true labels. The key lies in estimating this transition probability. Previous research(Chen and Gupta, 2015; Goldberger and Ben-Reuven, 2016) achieved this by adding a noise adaptation layer on top of the classification network and training it jointly. Later works(Patrini et al., 2017) estimate the transition probabilities independently, but this typically relies on noise-free validation data or additional assumptions.A third strategy for combating label noise is sample selection, which involves identifying and selecting clean samples from noisy data. For instance, Arpit et al.(Arpit et al., 2017) explored the tendency of deep networks to first learn simple (clean) patterns before gradually adapting to the memorization of noisy data. Based on this effect, Arazo et al. in 2019 used a bimodal Gaussian mixture model (GMM) to fit the distribution of sample losses, thereby distinguishing clean samples as those with lower losses. This method implies that networks can preferentially learn "cleaner" samples first when faced with complex, noisy data, providing another perspective for addressing the issue of noisy labels.In the robust training of multimodal models, Elad Amrani et al.(Amrani et al., 2021) proposed a noise estimation method based on multimodal density estimation. By leveraging the inherent correlations between different modalities, this method identifies noisy samples and improves the robustness of multimodal models, achieving comparable performance to state-of-the-art methods on multiple tasks. However, in the context of Large Language Models (LLMs), the method assumes that noisy labels arise from inconsistencies between modalities, which does not fully align with the situations of human or machine labeling errors. Furthermore, it does not sufficiently consider the interaction between noisy labels and the inherent knowledge of LLMs, as well as the importance of domain knowledge.

B The Label Noise Effects on VQA Tasks During CVIF

To comprehensively analyze the impact of label noise in both general and domain-specific contexts, we selected two widely recognized vision-based LLMs: Llava and MedVInT_TD (Zhang et al., 2023b). These models differ in their VIF processes but both involve training parameters θ and γ. To isolate the effects of label noise on θ and γ, we perform CVIF following the training procedure similar to Equation (2) but optimize different parts of the parameters independently. Initially, we freeze θ and fine-tune γ to solely investigate the influence of γ. Next, we freeze γ and fine-tune θ to observe the label noise effects on LLM's side. Finally, both θ and γ are fine-tuned simultaneously on the dataset to observe the overall influence under label noise. This case study allows for a detailed examination of how label noise distinctly affects each parameter.

The experimental results are presented in Table 2. It is evident that label noise in dataset $\mathcal{D}_C$ substantially impacts the learning of parameters γ and θ, leading to decreased accuracy in VQA tasks. This effect is observed in both open-ended and multiple-choice (cloze) VQA formats. Notably, when both γ and θ are influenced by label noise, the error is not just additive but compounded, leading to more pronounced inaccuracies. Therefore, in the development of the LEABNB framework, we focus on mitigating the impact of label noise from the perspectives of both γ and θ.

Table 2: Case study conducted on the SLAKE dataset

Method	0%	10%	20%	40%
SFT	74.45	70.53	65.96	45.42
Freeze Projection Layer	-	71.24	66.52	46.76
Freeze LLMs	-	68.83	63.25	42.66

C Experiment Details

C.1 Datasets

Experiments in this study were conducted on four open-source benchmark datasets: SLAKE (Liu et al., 2021a), VQA-RAD (Lau et al., 2018), GQA (Hudson and Manning, 2019), and OKVQA (Marino et al., 2019). The SLAKE and VQA-RAD datasets are dedicated to medical VQA tasks, while GQA and OKVQA are widely used for evaluating general VQA tasks. Since the aforementioned datasets are clean with accurately labeled data, to

simulate label noise, we employed a manual label corruption approach. Considering the resource-intensive nature of manual annotation, we opted to randomly select 3,000 samples from the training sets of the three datasets, excluding the complete training set of VQA-RAD, to construct our experimental training set. Additionally, we randomly drew 800 samples from each corresponding set to create our test set.

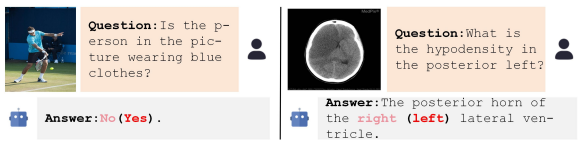

Figure 2: The figure illustrates the label noise discussed, with "closed-ended" and "open-ended" classes represented on the left and right, respectively. Red texts indicate clean labels, while pink texts denote manually annotated label noise. All manual noise perturbations are controlled within the same context, only changing target words' semantics without altering context, e.g., replacing "left" with "right" maintains context while introducing semantic noise.

As shown in Figure 2, we utilized a label flipping strategy for closed-ended question-answer pairs by altering "yes" labels to "no" with given probability and vice versa to generate erroneous labels, which was in alignment with previous works (Han et al., 2018; Liu et al., 2020; Xia et al., 2020). For open-ended question-answer pairs, we manually injected noise into the original labels with a given probability to simulate the presence of erroneous labels. This noise strategically introduces deviations in conceptual entities, effectively affecting label semantics without altering the overall semantic context of the sentence. This approach closely resembles actual noise scenarios commonly encountered in open-ended question-answering tasks, as incorrect labels often stem from subtle misunderstandings, ambiguities, or inherent biases present in the questions or answers. To ensure the robustness of our experiments, for datasets containing both types of question-answer pairs, we maintained randomness in our sampling and ensured a consistent ratio of closed to open-ended samples.

C.2 Vision-based LLMs, Hyperparameters, and Evaluation Metrics

We perform CVIF with two distinct models on two categories of datasets to assess our method in both general and domain-specific scenarios, with the different vision-based LLMs and hyperparameter configurations detailed as follows:

SLAKE and VQA-RAD: we employed the MedVInT-TD model (Zhang et al., 2023b), utilizing the AdamW optimizer with an initial universal learning rate set to 2×10^{-5}, without weight decay. The batch size was fixed at 8, and each experiment was conducted over five training epochs.

GQA and OKVQA: we deployed the LLaVA model (Liu et al., 2023) also using the AdamW optimizer. The initial learning rate was set at 2×10^{-5} for the LLM and 2×10^{-4} for the projection layer, both without weight decay. The experiments were conducted with a fixed batch size of 16 over five training epochs.

For closed-ended questions, we report accuracy as the performance metric. For open-ended questions, we employ recall to evaluate the proportion of true labels present within the generated sequences. For each task, we conduct five independent experiments with random seeds and report the mean accuracy as the result.

D Ablation

We conducted ablation studies on the GQA and OKVQA datasets to evaluate the effectiveness of the key components in our proposed methods, with results presented in Table 3. As discussed in Appendix B, overfitting to noisy labels deteriorates the performance of the projection layer and LLMs. To address this issue, we evaluate the effectiveness of two components: (1) the application of Polyak Averaging to the projection layer γ for gradual parameter updates, which reduces overfitting; and (2) the incorporation of Bias-learning in fine-tuning LLM θ to improve its robustness against noisy data. The experimental results demonstrate that these strategies significantly enhance the model's robustness in VQA tasks in the presence of label noise.

Table 3: Ablation results on GQA and OKVQA tasks

Task	Method	10%	20%	40%
GQA	Bias-learning	61.25	58.12	55.24
	Polyak Averaging	62.51	57.56	54.63
	LEABNB	**63.22**	**60.27**	**57.14**
OKVQA	Bias-learning	43.19	41.33	36.44
	Polyak Averaging	42.91	41.87	35.84
	LEABNB	**43.35**	**42.56**	**37.12**

To further elucidate the underlying mechanisms of bias-leaning, we conducted tests within a se-

mantic environment solely containing uncertainty prompts. The experimental results are detailed in Table 4. When relying exclusively on uncertainty prompts to guide the reasoning of MLLMs without allowing the model to learn from biases, the performance of the model was actually negatively impacted. These findings further underscore the importance and effectiveness of bias learning in handling noisy data.

Table 4: Results of uncertain inference under standard methods

Task	Method	10%	20%	40%
GQA	SFT	**60.84**	**54.63**	**51.85**
	Uncertain-inference	58.62	53.26	50.52
OKVQA	SFT	**42.76**	39.21	**34.74**
	Uncertain-inference	40.51	**39.86**	31.94