Learning LLM Preference over Intra-Dialogue Pairs: A Framework for Utterance-level Understandings

Xuanqing Liu*, **Luyang Kong***, **Wei Niu, Afshin Khashei, Belinda Zeng, Steve Johnson, Jon Jay, Davor Golac, Matt Pope**
Amazon.com Inc.

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in handling complex dialogue tasks without requiring use case-specific fine-tuning. However, analyzing live dialogues in real-time necessitates low-latency processing systems, making it impractical to deploy models with billions of parameters due to latency constraints. As a result, practitioners often prefer smaller models with millions of parameters, trained on high-quality, human-annotated datasets. Yet, curating such datasets is both time-consuming and costly. Consequently, there is a growing need to combine the scalability of LLM-generated labels with the precision of human annotations, enabling fine-tuned smaller models to achieve both higher speed and accuracy comparable to larger models. In this paper, we introduce a simple yet effective framework to address this challenge. Our approach is specifically designed for per-utterance classification problems, which encompass tasks such as intent detection, dialogue state tracking, and more. To mitigate the impact of labeling errors from LLMs – the primary source of inaccuracies in student models – we propose a noise-reduced preference learning loss. Experimental results demonstrate that our method significantly improves accuracy across utterance-level dialogue tasks, including sentiment detection (over 2%), dialogue act classification (over 1.5%), etc.

1 Introduction

Maintaining high annotation quality, scaling the size of labeled datasets, and managing annotation budgets are three critical yet often conflicting objectives in deploying real-world ML applications. A widely adopted paradigm involves a two-stage process: unsupervised pretraining followed by supervised fine-tuning (e.g., Devlin, 2018; Chen et al., 2020; He et al., 2020; Raffel et al., 2020). This approach effectively reduces the size of the labeled dataset required because, during the pretraining phase, models learn to generate universal embeddings across various modalities. Consequently, such pretrained models are often straightforward to adapt to downstream tasks.

In dialogue understanding, moving beyond BERT-like models is essential, as dialogues possess unique characteristics compared to the BERT pretraining corpus (which primarily consists of books and web pages). These differences arise from several factors: First, dialogues involve spoken language exchanges between two or more individuals and are often structured differently, with one line per speaker. This format reduces the effectiveness of tasks such as masked token prediction and next-sentence prediction. Second, the vocabulary in daily dialogues tends to be informal. Finally, dialogues are frequently transcribed from voice recordings, introducing ASR errors and background noise. These distinctive properties have inspired research into developing specialized unsupervised pretraining algorithms for dialogue data (Mehri et al., 2019; Zhong et al., 2022; Liu et al., 2022; Zhou et al., 2022). Benchmark evaluations on common dialogue tasks – such as intent detection, next-utterance prediction, summarization, dialogue act classification, and dialogue state tracking – demonstrate the advantages of dialogue-optimized models. These models generally adhere to the classical BERT framework, pretraining on large-scale unsupervised dialogue datasets with dialogue-specific loss functions, including random mask filling, utterance swapping, and contrastive learning. However, it remains unclear whether such pretrained embedding models generalize effectively to specific downstream tasks.

To address this challenge, we require direct supervision signals that are closely aligned with downstream tasks. This motivates the use of in-

*First two authors contributed equally. Corresponding author email: xuanqing@amazon.com

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 86–98

struction fine-tuned LLMs as phase-2 supervision signals, while retaining traditional unsupervised pretraining as phase-1. However, simply employing LLMs as data labelers and fine-tuning a student model using traditional cross-entropy loss proves suboptimal. The accuracy of LLM-generated labels can be unpredictable, influenced by factors such as the quality of the LLM, the prompting strategy, and the inherent difficulty of the dialogue task. Consequently, the knowledge transferred from the LLM to the student model often deviates from the intended objective. This paper proposes an alternative approach based on preference learning, where pairs of chunks sampled from the same dialogue session (*intra-session pairs*) are labeled by ensembled LLMs. Under reasonable assumption on LLM labeling errors, our method outperforms traditional training algorithms in both data efficiency and generalizability.

2 Related work

2.1 Task-oriented dialogue (TOD) system

Task-oriented dialogue understanding lies in the core of building AI assistants to be deployed in domain specific scenarios such as restaurant booking, self-service product troubleshooting, and so on. The objective is to help users achieve their goals in limited turns by understanding users' needs, tracking dialogue states and figure out next best action. Unique to TOD system, intent detection, dialogue act classification, and dialogue state tracking are three critical components of the system. Traditional approaches mostly rely on supervised learning on embedding models (Liu and Lane, 2016), by encoding dialogue contexts and employing deep neural networks such as RNN/LSTM or Transformers to infer utterance labels or slot values (Barriere et al., 2022; Duran, 2021; Chen et al., 2020). In the LLM age, there is a shift from finetuning TOD model for a specific domain (Lei et al., 2018) to open domain in-context learning (Hu et al., 2022; Arora et al., 2024). Unfortunately, both solutions ignored latency and cost constraints in real-time, commercial products.

2.2 Synthetic label prompting strategies and transfer learning

These two techniques are the foundation of our solution. We discuss the main idea and prior works.

Prompting strategies. It is often non-trivial prompting LLMs to achieve quality high data labeling. For example, prior work (Anagnostidis and Bulian, 2024; Work; Lu et al., 2021) noticed that few-shot prompting is surprisingly sensitive to factors including the number of example, order of examples, positive / negative sample ratio, or how similar those examples are to the actual input query. In this regard, fine-tuning embedding models on human curated labels are still preferred in production-ready applications. To strengthen the robustness of ICL, a promising solution is through diversified prompting (Li et al., 2023b; Song et al., 2024b,a), either by starting with a few seeding prompts, and augment more versions using automated pipeline (Wang et al., 2022b), or repetitively refine the prompt from diverse perspectives (Li et al., 2023a).

Transfer learning. For better instruction following ability, a popular approach is fine-tuning on synthetic datasets produced by larger LLMs (Taori et al., 2023; Chiang et al., 2023; Xu et al., 2023a). To foster LLM's reasoning ability, another line of work finetune with synthetic rationales collected from stronger LLMs (Wang et al., 2022a; Shridhar et al., 2023; Liu et al., 2023; Kang et al., 2024). Similar approach work for task-specific applications too, examples like dialogue generation (Xu et al., 2023b), information extraction (Josifoski et al., 2023; Jeronymo et al., 2023) and code generation (Chaudhary, 2023; Roziere et al., 2023). Our work focus on per-utterance multi-class classification in TOD system, assuming that even the most capable LLMs can't generate highly accurate labels, so a brand new transfer learning approach is required.

3 Proposed framework

3.1 Problem scope

We limit our scope to per-utterance classification, including sentiment detection, dialogue state tracking, dialogue act classification (Fig. 1).

Intent detection. Each utterance is mapped to a binary label `has_intent` ($y = 1$) or `no_intent` ($y = 0$). Positive label means utterance deemed a valid intent (e.g. a question, issue, or complaint). Take customer support for example, we could apply intent detection model to monitor customer speech in real time and figure out whether a customer is seeking for help rather than chit-chatting.

Dialogue act classification. We could regard this as an extension of intent detection from binary intent labels to multi-class acts. The objective of

(a) Intent detection

Utterances	Has intent?
[Assistant] Hi, this is [PII] speaking, how can I help you today?	No
[Customer] Hello, I have an issue with this security camera.	No
[Assistant] Okay?	No
[Customer] So, the green light shows it has connected to my phone.	No
[Customer] *which says no device found and so I couldn't see the recording.*	**Yes**
[Assistant] I do apologize to hear the problem. Let me find out the solution okay?	No

(b) Dialogue act classification

Utterances	Dialogue Act
[Doctor] Jackie, how are you?	Greeting
[Patient] Not too bad, how are you?	Greeting
[Doctor] Thanks for asking. What's going on there?	Information Request
[Patient] They think I have a drinking problem. My family ...	Information Delivery
[Doctor] Your family thinks you have a drinking problem?	Clarification Request
[Patient] Yeah. So we started this last weekend. They picked me up for my bridal shower. I drunk ...	Clarification Delivery

(c) Dialogue state tracking

Utterances	Dialogue State
[Assistant] Hi, this is XYZ hotel, how may I help?	N/A
[Customer] Hello, I want to book a room for Thanksgiving in San Francisco.	date: "Thanksgiving" city: "San Francisco"
[Assistant] Sure, happy to help. Any preference about the location? we have Bridge Garden at North San Francisco and the other one called Sonesta Inn close to the airport.	N/A
[Customer] Got it, we will stay in the north for 4 nights.	num_nights: 4 hotel: "Bridge Garden"
[Assistant] Sure! and do you have an account with us?	N/A

Figure 1: Illustrative examples of intent detection, dialogue act classification, and dialogue state tracking problems.

dialogue act classification is finding out the functions that utterances serve in dialogues – such as commitments, questions, requests, replies, etc. In contact centers, for example, classifying dialogue acts can be valuable at providing appropriate and thoughtful responses to clients adhering to the dialogue acts.

Dialogue state tracking (DST). The objective of DST is extracting and picking up new information into dialogue state as the conversation evolves. This task has great potential in customer service as it not only provides intent types (e.g. *hotel-booking* in Fig. 1c), but also identifies relevant semantic concepts throughout the slot filling process (e.g. *location* = *San Francisco*).

Challenge. When delivering real world applications driven by per-utterance classifiers, the challenges often rooted from obtaining high quality labels. For example, MultiWOZ (Budzianowski et al., 2018) is commonly used for benchmarking DST algorithms. Yet the original dataset contains numerous labeling errors, and it took 4 future versions (Eric et al., 2019; Zang et al., 2020; Han et al., 2021; Ye et al., 2021) (MultiWOZ 2.1-2.4) to correct them. More importantly, we learned that a clean dataset not only ensures us precisely tracking the progress on good valid/test set, but also reduces the reliance on robust model training algorithms (Ye et al., 2022). The challenge of labeling leads us to focus on following question –

> *Can we design a general solution for per-utterance classification problems, by jointly utilizing small amount of clean, human verified labels and almost unlimited amount of lower quality LLM annotations?*

We share a positive answer in the remainder of this work. Our work is not a simple extension of weakly supervised learning or noise-robust supervised learning, as we utilize characteristics that are unique to per-utterance classifications.

3.2 Workflow

Our workflow involves four stages. Goal of stage 1 is to construct a *prompt bank* containing diversified prompts that performs well on data annotation work following prompt tuning strategies outlined in Schulhoff et al. 2024; Brown et al. 2020; Wei et al. 2022; Yao et al. 2023; Liu et al. 2021. Predictions led by various prompts are slightly different, we ensemble the outputs together for better results (Khalifa et al., 2023; Jiang et al., 2021). Next, we further strengthen the ensemble effect at stage 2 using top-K/top-P sampling. After repeated sampling N times using LLM labeler, we compute L-dimensional score vector $S \in [0, 1]^L$ for dialogue $\mathcal{D}$ containing L utterances. Each element $0 \leq S_i \leq 1$ is the ratio of positive LLM labels divided by N (e.g. if 3 in 10 ensembles labeled i-th utterance as positive, $S_i = 0.3$). For C-class classification problem, we transform it into C one-versus-rest binary classification problems so the same framework still apply.

After we collect LLM labeling scores S, we split a dialogue into multiple segments using a sliding window of stride 1. We denote x_i as the i-th seg-

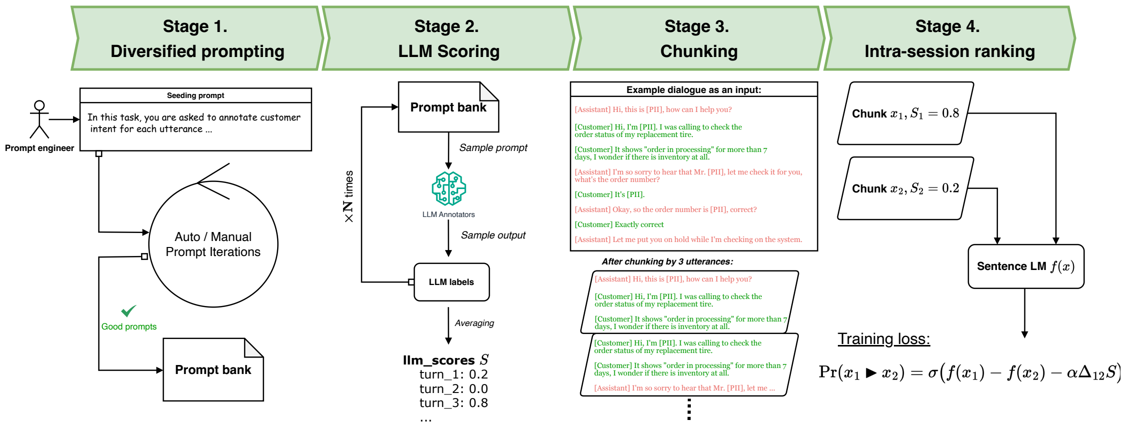

Figure 2: Overview of our framework to train a small student model using noisy LLM supervision.

ment covering u_1 to u_i. Finally in stage 4, we randomly sample two *intra-session* segments x_i and x_j from the same dialogue and train a student model f minimizing pair-wise ranking loss:

$$\ell(x_i, x_j) = \mathrm{KL}\big(\mathbb{I}_{y_i \blacktriangleright y_j} \parallel \Pr(x_i \blacktriangleright x_j)\big), \quad (1)$$

where $\mathbb{I}_{y_i \blacktriangleright y_j} = 1$ iff. $y_i = 1$ and $y_j = 0$ for binary labels; $\Pr(x_i \blacktriangleright x_j)$ is the probability of x_i being more positive than x_j, modeled by network f under an adaptive margin:

$$\Pr(x_i \blacktriangleright x_j) = \sigma\big(\Delta_{i,j} f - \alpha \cdot \Delta_{i,j} S\big), \quad (2)$$

where σ is the Sigmoid function, $\Delta_{i,j} f = f(x_i) - f(x_j)$ is the difference of model predicted scores and $\Delta_{i,j} S = S_i - S_j$ is the difference of LLM predicted scores between segment i and j; $\alpha \in [0, 1]$ is a tunable hyper-parameter controlling margin. We train a student network f over intra-session pairs to ensure: for any positive+negative pair labeled by LLM (positive x_i *vs.* negative x_j), the student network f has the same preference as teacher LLM under margin $\alpha \cdot \Delta_{i,j} S$. This idea made two hidden assumptions: First assuming the LLM score S is a good estimator of ground-truth correctness probability (*aka.* confidence calibrated (Guo et al., 2017)); secondly, single LLM labeler may be biased and high variance, their difference within same dialogue session $S_i - S_j$ carries dramatically lower bias and variance due to the differentiation. Therefore estimation error of $S_i - S_j$ is more precise than S_i or S_j alone. We discuss and verify two assumptions in the following sections.

3.3 Stage 1-2: How well are LLM scores calibrated to accuracy?

A desirable property of LLM teacher is confidence scores S calibrated to labeling accuracy, i.e. we expect higher true-positive rate if LLM score S_i closes to one; and near zero true-positive rate if S_i is closer to zero:

$$\Pr(y_i = 1 | S_i) = S_i. \quad (3)$$

If Eq. (3) is true, we could replace ground truth label y_i with soft label S_i without incurring additional gradient bias and variance (see Appendix F for a proof). In addition, Eq. (3) implies monotonicity relationship:

$$S_i > S_j \implies \Pr(y_i = 1) > \Pr(y_j = 1). \quad (4)$$

(Guo et al., 2017) showed that DNNs are uncalibrated, in that their accuracy falls behind confidence score (DNNs are over-confident). Same findings are reported in LLM world (Kapoor et al., 2024; Huang et al., 2024). Among various post-training solutions to calibrate DNNs (e.g. (Zadrozny and Elkan, 2001; Mozafari et al., 2018)), one simple and effective technique is ensemble different models (Lakshminarayanan et al., 2017) which integrates well with our workflow. Remaining question to be answered in this work is -

> *Does the same ensemble technique work for LLM predictions? If so, how many ensemble predictions we need to calibrate the scores?*

We design following experiment to answer this question: We sample an intent detection dataset containing around 600 transcripts and binary `has_intent` / `no_intent` per-utterance labels. A labeling prompt optimized for `Claude3-sonnet`[1] for this task is provided in Appendix E. We apply the same prompt to ensemble sizes n between 1 and 30. In each setting, we run LLM labeling on each input pair $\langle x_i, x_j \rangle$ for n times and obtain scores S_i and S_j by averaging LLM predictions. Lastly, we partition the data by value S_i into five buckets: $S_i \in (0.0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, $(0.8, 1.0]$. Within each bucket, we compute the percentage of positive ground-truth labels. We apply ECE loss, the standard metric to measure DNN calibration error (Guo et al., 2017):

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left| \text{acc}(B_m) - \text{conf}(B_m) \right| \quad (5)$$

where B_m is the m-th bucket partitioned by S_i. $\text{acc}(B_m) = \Pr(y_i = 1 | s_i \in B_m)$ is the accuracy of B_m; and $\text{conf}(B_m)$ is the overall confidence score in B_m. Due to Eq. (3) lower ECE metric means better calibration. Despite some random fluctuations, we could observe in Fig. 3 a decline in ECE loss (0.22 $\searrow$ 0.17) as ensemble size increases.

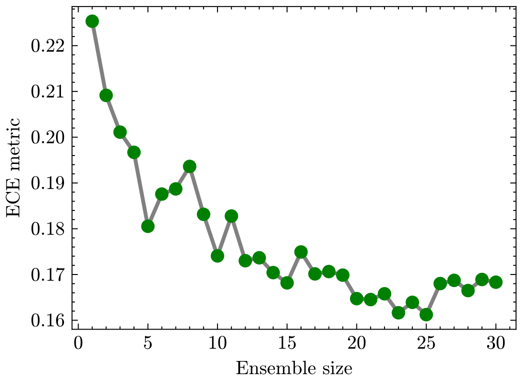

Figure 3: Visualizing the downward trend of ECE loss as ensemble size increases from 1 to 30.

> *The ensemble technique in Stage 1-2 effectively calibrates LLM scores S_i by introducing fewer gradient biases and variances. Therefore LLM teacher supervisions are good surrogate for ground-truth labels.*

3.4 Stage 3-4: Overcoming distribution shifts by intra-session comparison

We generate ranking pairs in a novel way: we sample two chunks for ranking from the same conversation (*intra-session pairs*), instead of different conversations. We make two hypothesis (H_1 and H_2) explaining why intra-session pairs are more powerful.

H_1**: Intra-session pairs are harder**. Two chunks sampled from same dialogue are similar in the context (sharing the same topic with overlapping context). As a result, it is harder to tell which chunk is positive label against the other. Once training a student model on top of hard pairs, it forces the model to learn more discriminative textual features from text input, rather than just replying on some keywords. Those intra-session pairs lead to better generalization.

H_2**: LLM labeling errors are canceled by the differentiator**. This hypothesis is more conceptually involved: LLM labeling errors are not uniformly random across all data, instead they cluster on certain type of transcripts. For example, some scenarios are not mentioned in the labeling prompt so LLM has to guess, resulting in more errors in such cases. Fortunately, this type of error typically condensed to certain dialogues, equivalent to a "shifting" effect to the label distribution. By sampling a pair (x_i and x_j) from the same dialogue, their corresponding LLM scores (S_i and S_j) are drifted to roughly the same extent. In the end, the margin of the loss function (1) $\Delta_{ij} S = S_i - S_j$ still accurately tracking ground-truth label difference $y_i - y_j$.

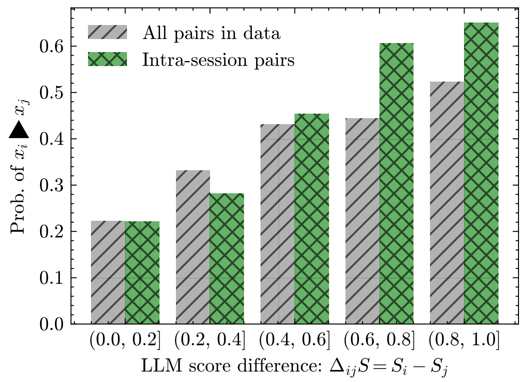

Figure 4: Comparing the correlations between LLM score difference (also the margin of training loss) *w.r.t.* the probability of one label is more positive than the other. We also include linear fittings to both groups.

We design an experiment to validate H_2 on two groups: the control group consists of pairs sampled from different dialogues; experimental group consists of pairs sampled from same dialogue. The goal is checking correlation between $\Delta_{ij} S = S_i - S_j$ with the probability of $y_i = 1$ and

[1]Available at Anthropic and AWS Bedrock.

$y_j = 0$ ($y_i > y_j$ in binary case). We follow the same bucketizing method as previous experiment (5 buckets). We count the percent of $y_i > y_j$ cases in each bucket and each group. Result in Fig. 4 shows the ground-truth probability of $y_i > y_j$ more sensitive to $\Delta_{ij}S$ in experimental group than control group. Meaning that our intra-session pairs are indeed less noisy, and a better approximation of golden supervision signal $y_i - y_j$.

4 Experiments

Datasets. We benchmark our method on three important tasks in task-oriented dialogues (TOD): intent/sentiment-detection, dialogue act classification, and dialogue state tracking. We benchmark intent/sentiment detection on MELD (Poria et al., 2019) and SILICONE (Busso et al.); benchmark dialogue act classification on daily-dialog (Li et al., 2017), MRDA (Shriberg et al., 2004), BT-OASIS (Duran, 2021) and dyda_da (Chapuis et al., 2020); benchmark dialogue state tracking on SGD (Rastogi et al., 2020) and MultiWOZ-2.2 (Zang et al., 2020). We put statistics and other details of datasets in Appendix A.

Baselines. We want to see how the accuracy change after plugging our workflow into some strong models. We select following baselines accordingly:

- *Claude3-Sonnet*: We pick this model as a strong baseline for measuring LLM annotator performance.
- *FnCTOD* (Li et al., 2024): A recent prompting strategy achieving strong results on dialogue state tracking task.
- *ToD-BERT* (Wu et al., 2020): A strong baseline for dialogue pretrained small embedding model. This is also the backbone model of our method.
- *FLAN-T5* (Chung et al., 2024): T5-XXL fine-tuned on large-scale instructions data including MultiWOZ. We include this model as a natural baseline for fine-tuned LLM on TOD datasets.

We summarize features of all baselines with our method in Table 6 of Appendix B.

4.1 Comparing pairwise preference learning *vs.* pointwise knowledge transfer

To evaluate the transition from pointwise model distillation to pairwise preference learning, we compare the intent detection accuracy of the ToD-BERT model fine-tuned using three approaches: 1) fine-tuning directly on human-labeled data; 2) supervised pretraining with pointwise LLM-generated labels followed by fine-tuning on human-labeled data; and 3) supervised pretraining with pairwise LLM-generated labels followed by fine-tuning on human-labeled data. To assess the impact of data scaling, we vary the sampling ratios during evaluation. Table 1 consistently shows that models leveraging pairwise supervised pretraining outperform the alternatives, particularly in low-data regimes.

Approach \ % gold labels	0%	1%	5%	10%	25%
Finetune-only	-	27.3	29.5	34.7	69.6
Supervised pretrain → Finetune					
Pointwise pretrain	-	31.8	33.4	47.2	77.3
Pairwise pretrain	-	38.4	45.8	52.1	78.4

Table 1: Effective of our approach under various amount of labeled data.

4.2 Sentiment detection

Next we benchmark our method with baselines on two sentiment detection datasets. We report classification accuracy over all sentiments defined in each datasets. The results are shown in Table 2. Comparing with ToD-BERT (finetuned directly on human labeled data) and FnCTOD (finetuned on LLM synthetic data), our approach (supervised pretrained on LLM synthetic data using pairwise loss then finetuned on human labeled data) performs better than baselines by around 2% to 8%.

Datasets	Claude	FnCTOD	ToD-BERT	FLAN-T5	**Ours**
MELD	74.25	68.84	80.30	75.72	88.09
IEMOCAP	76.39	61.30	87.88	82.62	90.31

Table 2: Benchmarking intent/sentiment detection task.

4.3 Dialogue act classification

Similarly, we benchmark our method against baselines on dialogue act classification problem. Note we adopted the same backbone model as ToD-BERT, and ToD-BERT is still the strongest baseline in this task. Our model out-performed ToD-BERT by around 1.5% to 10%.

Datasets	Claude	FnCTOD	ToD-BERT	FLAN-T5	**Ours**
DailyDialog	70.39	66.03	72.40	68.08	76.50
MRDA	62.82	81.93	88.4	60.47	89.95
dyda_da	71.25	74.82	79.14	68.66	85.11
BT-Oasis	32.85	52.76	59.24	17.13	69.62

Table 3: Benchmarking dialogue act classification task.

4.4 Dialogue state tracking

Finally, we benchmark on two dialogue state tracking (DST) datasets, SGD and MultiWOZ-2.1. In this experiment we benchmark the accuracy of joint prediction of slot/domain/values (aka. **Joint-Acc**). The results are shown in Figure 4.

Datasets	Claude	FnCTOD	ToD-BERT	FLAN-T5	**Ours**
SGD	60.7	63.9	42.5	–	47.3
MultiWOZ	27.0	37.9	16.4	–	25.5

Table 4: Benchmarking dialogue state tracking task.

5 Discussion and future work

This paper presents a novel approach to minimizing human effort in labeling high-quality data for a class of per-utterance classification problems. Our method moves beyond traditional LLM labeling and knowledge transfer to student models by leveraging a preference learning and pairwise ranking framework. This framework has been demonstrated to be both theoretically and empirically robust against LLM labeling errors. An intriguing future direction would be to extend this approach to reward model training in reinforcement learning with human feedback (RLHF), another critical domain characterized by noisy labels and the need for robust discriminative model training.

References

Sotiris Anagnostidis and Jannis Bulian. 2024. How susceptible are llms to influence in prompts? *arXiv preprint arXiv:2408.11865*.

Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. Intent detection in the age of llms. *arXiv preprint arXiv:2410.01627*.

Valentin Barriere, Slim Essid, and Chloé Clavel. 2022. Opinions in interactions : New annotations of the SEMAINE database. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7049–7055, Marseille, France. European Language Resources Association.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

C Busso, M Bulut, CC Lee, A Kazemzadeh, E Mower, S Kim, JN Chang, S Lee, and SS Narayanan IEMOCAP. Interactive emotional dyadic motion capture database., 2008, 42. *DOI: https://doi.org/10.1007/s10579-008-9076-6*, pages 335–359.

Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloé Clavel. 2020. Hierarchical pre-training for sequence labelling in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2636–2648, Online. Association for Computational Linguistics.

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation. *Code alpaca: An instruction-following llama model for code generation*.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

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)*, 2(3):6.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Nathan Duran. 2021. Bt-oasis corpus.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Ting Han, Ximing Liu, Ryuichi Takanabu, Yixin Lian, Chongxuan Huang, Dazhen Wan, Wei Peng, and Minlie Huang. 2021. Multiwoz 2.3: A multi-domain task-oriented dialogue dataset enhanced with annotation corrections and co-reference annotation. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part II 10*, pages 206–218. Springer.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah A Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. *arXiv preprint arXiv:2203.08568*.

Yukun Huang, Yixin Liu, Raghuveer Thirukovalluru, Arman Cohan, and Bhuwan Dhingra. 2024. Calibrating long-form generations from large language models. *arXiv preprint arXiv:2402.06544*.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

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.

Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574.

Minki Kang, Seanie Lee, Jinheon Baek, Kenji Kawaguchi, and Sung Ju Hwang. 2024. Knowledge-augmented reasoning distillation for small language models in knowledge-intensive tasks. *Advances in Neural Information Processing Systems*, 36.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Exploring demonstration ensembling for in-context learning. *Preprint*, arXiv:2308.08780.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Dawei Li, Yaxuan Li, Dheeraj Mekala, Shuyao Li, Xueqi Wang, William Hogan, Jingbo Shang, et al. 2023a. Dail: Data augmentation for in-context learning via self-paraphrase. *arXiv preprint arXiv:2311.03319*.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Omer Levy, Luke Zettlemoyer, Jason E Weston, and Mike Lewis. 2023b. Self-alignment with instruction backtranslation. In *The Twelfth International Conference on Learning Representations*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zekun Li, Zhiyu Zoey Chen, Mike Ross, Patrick Huber, Seungwhan Moon, Zhaojiang Lin, Xin Luna Dong, Adithya Sagar, Xifeng Yan, and Paul A Crook. 2024. Large language models as zero-shot dialogue state tracker through function calling. *arXiv preprint arXiv:2402.10466*.

Bing Liu and Ian Lane. 2016. Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Che Liu, Rui Wang, Junfeng Jiang, Yongbin Li, and Fei Huang. 2022. Dial2vec: Self-guided contrastive learning of unsupervised dialogue embeddings. *arXiv preprint arXiv:2210.15332*.

Hanmeng Liu, Zhiyang Teng, Leyang Cui, Chaoli Zhang, Qiji Zhou, and Yue Zhang. 2023. Logicot: Logical chain-of-thought instruction tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? *arXiv preprint arXiv:2101.06804*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2021. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*.

Shikib Mehri, Evgeniia Razumovskaia, Tiancheng Zhao, and Maxine Eskenazi. 2019. Pretraining methods for dialog context representation learning. *arXiv preprint arXiv:1906.00414*.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2018. Attended temperature scaling: a practical approach for calibrating deep neural networks. *arXiv preprint arXiv:1810.11586*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Tal Remez, Jérémy Rapin, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, et al. 2024. The prompt report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 97–100.

Kumar Shridhar, Alessandro Stolfo, and Mrinmaya Sachan. 2023. Distilling reasoning capabilities into smaller language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7059–7073.

Feifan Song, Bowen Yu, Hao Lang, Haiyang Yu, Fei Huang, Houfeng Wang, and Yongbin Li. 2024a. Scaling data diversity for fine-tuning language models in human alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14358–14369.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024b. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

PeiFeng Wang, Aaron Chan, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022a. Pinto: Faithful language reasoning using prompt-generated rationales. In *The Eleventh International Conference on Learning Representations*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

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

What Makes In-Context Learning Work. Rethinking the role of demonstrations: What makes in-context learning work?

Chien-Sheng Wu, Steven Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. *arXiv preprint arXiv:2004.06871*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6268–6278.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Preprint*, arXiv:2305.10601.

Fanghua Ye, Yue Feng, and Emine Yilmaz. 2022. Assist: Towards label noise-robust dialogue state tracking. *arXiv preprint arXiv:2202.13024*.

Fanghua Ye, Jarana Manotumruksa, and Emine Yilmaz. 2021. Multiwoz 2.4: A multi-domain task-oriented dialogue dataset with essential annotation corrections to improve state tracking evaluation. *arXiv preprint arXiv:2104.00773*.

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, volume 1, pages 609–616.

Xiaoxue Zang, Abhinav Rastogi, Srinivas Sunkara, Raghav Gupta, Jianguo Zhang, and Jindong Chen. 2020. Multiwoz 2.2: A dialogue dataset with additional annotation corrections and state tracking baselines. *arXiv preprint arXiv:2007.12720*.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Zhihan Zhou, Dejiao Zhang, Wei Xiao, Nicholas Dingwall, Xiaofei Ma, Andrew O Arnold, and Bing Xiang. 2022. Learning dialogue representations from consecutive utterances. *arXiv preprint arXiv:2205.13568*.

A Summary statistics of experiment datasets

Data	#Classes	#Dialogues	#Utterances
Intent/Sentiment detection			
MELD	3	1,400	13,000
IEMOCAP	6	151	10,039
Dialogue act classification			
DailyDialog	5	13,118	103,630
MRDA	5	75	108,202
dyda_da	4	87170	102,000
BT-Oasis	42	636	15,067
Dialogue state tracking			
SGD	53 (slots)	16,142	329,964
MultiWOZ-2.1	24 (slots)	8,438	42,190

Table 5: Datasets for each evaluation task and some statistics.

B Comparing features of baseline models and our method

Methods	TOD finetuned?	LLM distilled	Small size
Claude	(unknown)	✗	✗
FnCTOD	✗	✔	✗
ToD-BERT	✔	✗	✔
FLAN-T5	✔	✗	✗
Ours	✔	✔	✔

Table 6: Comparing baselines and our method along three dimension: TOD finetuned means whether the model is finetuned for TOD tasks; LLM distilled indicates the model is distilled from (imperfect) LLM synthetic labels; Small size means whether the actual inference model is small footprint.

C Sample prompts for Claude

Prompt for daily-dialogue:

```
Dialogue:
{dialogue}

Last utterance:
{last_utterance}

What's the best dialogue act of the last
utterance?
Choose from below without further
explain:

Options:
A. Inform
B. Question
C. Directive
D. Commissive
E. None of above

A valid output should be one of: A, B, C,
D, or E
Do not output anything else.
```

Prompt for MRDA:

```
Dialogue:
{dialogue}

Last utterance:
{last_utterance}

What's the best dialogue act of the last
utterance? Choose from below without
further explain:

Options:
A. Statement or subjective statement
B. Declarative question
C. Backchannel
D. Follow-me
E. Question

A valid output should be one of: A, B, C,
D, or E
Do not output anything else.
```

Prompt for MELD:

```
## Task Description

In this task you will receive a short
dialogue. Your goal is to read the whole
dialogue, understand the sentiment of
each utterances, and pick out the utter-
ances with positive sentiment.

## Output format

You need to copy each positive sentiment
utterances to an json array together
with the initial line number.

## Example

Input:

1 [Phoebe] Oh my God, he's lost it. He's
totally lost it.
2 [Monica] What?
3 [Ross] Or! Or, we could go to the bank,
close our accounts and cut them off at
the source.
4 [Chandler] You're a genius!
5 [Joey] Aww, man, now we won't be bank
buddies!
6 [Chandler] Now, there's two reasons.
7 [Phoebe] Hey.
8 [All] Hey!
9 [Phoebe] Ohh, you guys, remember that
cute client I told you about? I bit him.
10 [Rachel] Where?!
11 [Phoebe] On the touchy.

Correct output:
```json
{
    "positive_utterances": [
        "4 [Chandler] You're a genius!",
        "8 [All] Hey!"
    ]
}
```
```

```
‘‘‘
```

D Sample prompts for FLAN-T5

Prompt for daily-dialogue:

```
Dialogue:
{dialogue}

Last utterance:
{last_utterance}

What's the best dialogue act of the last
 utterance?

Options:
A. Inform
B. Question
C. Directive
D. Commissive
E. None of above
```

Prompt for MRDA:

```
Dialogue:
{dialogue}

Last utterance:
{last_utterance}

What's the best dialogue act of the last
utterance? Choose from below without
further explain:

Options:
A. Statement or subjective statement
B. Declarative question
C. Backchannel
D. Follow-me
E. Question

Answer:
```

Prompt for MELD:

```
Dialogue:
{dialogue}

Last utterance:
{last_utterance}

Is the last utterance in positive
sentiment? Choose "Yes" or "No".
```

E Intent detection labeling prompt

```
# Task description
You are given a conversation between user
and assistant. Typically, the user has
some questions / issues / complaints.
Your goal is to find out the utterance
containing the user intent.

# Data description
Each line of the conversation corresponds
to an utterance. You can see the speaker
from according to the beginning of each
line. For example:
```

```
‘‘‘
[assistant] Hi, my name is [PII], thank
you for calling [COMPANY].
[user] Hi, I'm calling because the
shippment arrived damaged and I need a
replacement.
[assistant] I see, I'm sorry to hear
your bad experience about shippment.
‘‘‘

Here the user intent is "Hi, I'm calling
because the shippment arrived damaged
and I need a replacement.".

Now it is your turn, read the
conversation thoroughly and find out all
intent utterances

Conversation:
{conversation}
```

F Proof of Unbiased Gradients

Theorem 1. *Suppose dataset $\{(x_i, y_i)\}$ has binary labels $y_i \in \{0, 1\}$. If we only have access to noise-corrupted soft labels $\{x_i, \hat{y}_i\}$, $\hat{y}_i \in [0, 1]$ where the noisy labels follow the property $\Pr(y_i = 1|\hat{y}_i) = \hat{y}_i$ (perfect confidence calibration). Then if we train a linear classifier $f_\theta(x) = \sigma(\theta^T x)$ on corrupted dataset the gradients of cross-entropy loss over parameters θ are unbiased.*

Proof. Training on corrupted dataset $\{x_i, \hat{y}_i\}$ using cross-entropy loss with linear model, we have the loss function:

$$L\big(\theta; (x_i, \hat{y}_i)\big) = -\hat{y}_i \log\big(f_\theta(x_i)\big) - (1 - \hat{y}_i)\log\big(1 - f_\theta(x_i)\big) \tag{6}$$

If we compute the gradients of loss over parameters θ:

$$\frac{\partial}{\partial \theta} L\big(\theta; (x_i, \hat{y}_i)\big) = \big(f_\theta(x_i) - \hat{y}_i\big)x_i. \tag{7}$$

If we take the expectation over randomness of $\hat{y}_i$ on both sides of Eq. (7), we can further get

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} L\big(\theta; (x_i, \hat{y}_i)\big)\right] = \big(f_\theta(x_i) - \mathbb{E}[\hat{y}_i]\big)x_i. \tag{8}$$

Furthermore, due to the calibration of $\hat{y}_i$, $\Pr(y_i = 1|\hat{y}_i) = \hat{y}_i$, we have that

$$\hat{y}_i = \Pr(y_i = 1|\hat{y}_i) = \mathbb{E}[y_i|\hat{y}_i]. \tag{9}$$

Taking expectation on both sides in Eq. (9), and leveraging the low of total expectation, we get

$$\mathbb{E}[\hat{y}_i] = \mathbb{E}[\mathbb{E}[y_i|\hat{y}_i]] = \mathbb{E}[y_i]. \tag{10}$$

Finally, we plug Eq. (10) into Eq. (8):

$$\begin{aligned}
&\mathbb{E}\left[\frac{\partial}{\partial\theta}L(\theta;(x_i,\hat{y}_i))\right]\\
&=\big(f_\theta(x_i)-\mathbb{E}[\hat{y}_i]\big)x_i\\
&=\big(f_\theta(x_i)-\mathbb{E}[y_i]\big)x_i\\
&\mathbb{E}\left[\frac{\partial}{\partial\theta}L(\theta;(x_i,y_i))\right].
\end{aligned} \tag{11}$$

Therefore we have proved that well-calibrated training dataset $\{x_i,\hat{y}_i\}$ is unbiased training of the model. □