# A Controlled Study of Auxiliary Supervision for Low-Resource Emotion Classification

**Martina Szabóová and Daniel Zemančík and Ján Magyar**
Institute of Artificial Intelligence, Technical University of Košice, Košice, Slovakia
martina.szaboova@tuke.sk

## Abstract

We present a controlled empirical study of auxiliary supervision for low-resource emotion classification. Using a shared-encoder multi-task learning framework, we compare three auxiliary objectives—coarse affective valence, valence–activation structure, and positive-affect detection—across three supervision sources: aligned labels deterministically derived from gold emotion annotations, rule-based weak labels generated with VADER, and synthetic labels generated by Qwen2.5-14B. We evaluate on MELD, GoEmotions, and DAIR-AI Emotion under matched training procedures and multiple data regimes. Auxiliary supervision yields its clearest, though modest, gains in low-resource settings, especially on MELD and GoEmotions, while improvements on DAIR-AI Emotion are smaller and less consistent. Error and per-class analyses show that the largest effects—both improvements and degradations—occur among low-frequency or semantically overlapping categories such as *anger*/*disgust*, *fear*/*sadness*, *remorse*, *embarrassment*, *pride*, and *grief*. We further find that agreement with aligned auxiliary labels is only weakly associated with downstream transfer, suggesting that useful auxiliary supervision need not exactly reproduce gold-derived affective mappings. Overall, our results support auxiliary supervision as a lightweight inductive bias for low-resource emotion classification, while also showing that its benefits are dataset-dependent and sensitive to auxiliary-task design.

## 1 Introduction

Despite major advances in pretrained language models, fine-grained emotion classification remains difficult, especially when labeled training data are limited. Emotion labels often have fuzzy semantic boundaries, and related categories such as *anger* and *disgust* or *fear* and *sadness* are frequently confused (Demszky et al., 2020; Barnes, 2023; Gong et al., 2025). These challenges are amplified in conversational emotion recognition, where emotional interpretation may depend on dialogue context, speaker interaction, and pragmatic cues (Poria et al., 2019; Majumder et al., 2019; Ghosal et al., 2020).

Multi-task learning (MTL) offers a natural way to address these challenges by training a primary task jointly with related auxiliary objectives. When the auxiliary task captures useful structure, MTL can reduce overfitting and encourage transferable representations, particularly in low-resource settings (Caruana, 1997; Chen et al., 2024; Crawshaw, 2020). In emotion recognition, prior work has explored auxiliary objectives such as sentiment prediction, emotion-shift prediction, and dialogue-level affective dynamics (Gao et al., 2022; Wang and Mine, 2023). However, comparatively little is known about how the *source* of auxiliary supervision affects transfer when the auxiliary labels differ in fidelity, noise, and abstraction.

This issue is increasingly important because large language models (LLMs) are now widely used to generate or assist annotations for affective NLP. Recent work on LLM-assisted emotion annotation and synthetic supervision shows that LLM-generated labels can be useful, but also that they may diverge systematically from human or dataset-specific labels, especially for ambiguous or negative emotions (Tan et al., 2024; Pangakis and Wolken, 2024; Niu et al., 2024; Greschner and Klinger, 2025; Christ and Amiriparian, 2026). We therefore examine whether auxiliary supervision helps by faithfully reproducing gold-derived labels or by imposing broader affective structure that remains useful even when labels are noisy.

We study this question through a controlled comparison of auxiliary supervision sources for emotion classification. Using a shared-encoder MTL architecture, we compare three auxiliary objectives—coarse affective valence, valence–activation struc-



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 24407–24420
October 24-29, 2026 ©2026 Association for Computational Linguistics

ture, and positive-affect detection—under three supervision sources: aligned gold-derived labels, rule-based weak labels from VADER, and LLM-generated labels from Qwen2.5-14B. We evaluate these configurations on MELD, GoEmotions, and DAIR-AI Emotion across multiple training-data fractions.

Our goal is analytical rather than architectural: we examine when auxiliary supervision helps, whether higher auxiliary-label agreement predicts stronger transfer, and which emotion categories are most affected.

Our contributions are as follows:

- We present a controlled comparison of aligned, rule-based, and LLM-generated auxiliary supervision for emotion classification within a unified MTL framework.
- We show that auxiliary supervision is most beneficial in low-resource settings, with the clearest gains on MELD and GoEmotions.
- We analyze the relationship between auxiliary-label agreement and downstream performance, finding that higher agreement does not necessarily correspond to larger transfer gains.
- We provide confusion-based and per-class analyses showing that the largest effects of auxiliary supervision—both improvements and degradations—occur among low-frequency or semantically overlapping emotion categories.

## 2 Related Work

### 2.1 Emotion Classification and Conversational Emotion Recognition

Pretrained transformer encoders substantially improved text classification and emotion recognition (Devlin et al., 2019; Liu et al., 2019), but fine-grained affective labels remain difficult because categories are semantically overlapping, subjective, and often imbalanced (Demszky et al., 2020; Gong et al., 2025); linguistic-feature enrichment has also been explored (Lecourt et al., 2026). Conversational emotion recognition further complicates the task because utterance-level emotions may depend on dialogue context, speaker state, and discourse structure (Poria et al., 2019; Majumder et al., 2019; Ghosal et al., 2020; Gao et al., 2024; Fu et al., 2025), and minority or overlapping categories remain unstable in low-resource settings. Perspectivist and disagreement-aware work further argues that subjective emotion labels can reflect valid annotator differences rather than pure noise (Fornaciari et al., 2021; Uma et al., 2021; Mostafazadeh Davani et al., 2022), motivating caution when interpreting both gold and automatically generated auxiliary labels.

### 2.2 Multi-Task Learning and Auxiliary Supervision

Multi-task learning improves generalization by jointly optimizing related objectives and sharing representations (Caruana, 1997; Chen et al., 2024; Crawshaw, 2020), but transfer is not guaranteed: gains depend on task relatedness, loss weighting, label quality, and gradient dynamics, and poorly aligned tasks can cause negative transfer (Standley et al., 2020; Chen et al., 2018; Yu et al., 2020; Qin et al., 2025; Borsani et al., 2026).

In emotion recognition, auxiliary tasks have often been based on coarse sentiment, emotion shift, or dialogue-level affective structure (Gao et al., 2022; Wang and Mine, 2023), but these studies typically do not isolate the effect of the supervision source under a fixed auxiliary taxonomy—the gap our work addresses by comparing aligned, weak, and LLM-generated auxiliary labels while holding the primary architecture, auxiliary tasks, and training procedure fixed.

### 2.3 Weak and Noisy Supervision

Weak supervision reduces annotation cost via heuristics, lexicons, or model predictions (Ratner et al., 2017), and VADER remains a common lightweight source of rule-based affective supervision (Hutto and Gilbert, 2014). Pseudo-labeling and noisy-label learning further show that imperfect labels can still improve downstream models (Arazo et al., 2020; Xie et al., 2020; Song et al., 2023). Most of this work, however, treats approximate labels as replacements for gold labels; comparatively little attention has been paid to how noisy affective labels behave when used only as auxiliary supervision in MTL, where a label may be useful even if it is not reliable enough to replace the primary emotion label.

### 2.4 LLM-Generated Synthetic Supervision

LLMs are increasingly used as annotators in NLP (Tan et al., 2024), and LLM-generated labels can

support competitive downstream classifiers, including in emotion-related settings (Pangakis and Wolken, 2024; Rahman and Caragea, 2025; Niu et al., 2024; Christ and Amiriparian, 2026). However, LLM-generated emotion labels can diverge from human labels in systematic ways—for example by overpredicting negative emotions or behaving inconsistently across prompt formulations (Greschner and Klinger, 2025). Rather than assuming that LLM labels are faithful replacements for human annotation, we test whether they can still provide useful auxiliary structure when optimized jointly with gold primary labels.

## 3 Methodology

We study how auxiliary supervision source affects transfer performance in emotion classification. To isolate this factor, all experiments use the same shared-encoder MTL architecture and training procedure while varying only the auxiliary-task formulation and the source of auxiliary labels.

### 3.1 Primary Task

The primary task, denoted Task A, is emotion classification. For MELD and DAIR-AI Emotion, this is single-label multiclass classification with 7 and 6 emotion categories, respectively. For GoEmotions, this is multi-label classification over 27 emotion categories plus *neutral*, for 28 labels in total.

### 3.2 Auxiliary Objectives

We define three auxiliary objectives that abstract the original emotion labels into coarser affective categories:

- **Task B: Coarse affective valence.** Emotions are mapped into *positive*, *negative*, or *ambiguous* categories.
- **Task C: Valence–activation structure.** Negative affect is divided into *negative_active* and *negative_passive*, while positive and ambiguous categories are retained.
- **Task D: Positive-affect detection.** The auxiliary target is reduced to a binary distinction between *positive* and *non_positive*.

Motivated by valence and activation dimensions in Russell's circumplex model of affect (Russell, 1980), these objectives form a controlled granularity ladder: Task B captures valence, Task C adds an active/passive split of negative affect, and Task D reduces the target to a positive vs. non-positive binary. We selected these abstractions because they can be applied consistently across the evaluated datasets through dataset-specific mappings. We do not treat them as complete psychological models of emotion; instead, they provide controlled auxiliary abstractions for studying how task granularity and supervision quality influence transfer.

Each auxiliary objective is implemented with a separate task-specific classification head attached to the shared encoder. During training, the model optimizes the primary emotion objective together with one auxiliary objective at a time. We do not jointly optimize multiple auxiliary objectives in the same model. Because the auxiliary taxonomy is held fixed within each task and only the supervision source (aligned, VADER-weak, or LLM) varies, differences observed within a given auxiliary task can be attributed to variation in the supervision source rather than to the taxonomy itself.

For MELD and DAIR-AI Emotion, the joint loss is:

$$\mathcal{L} = \mathcal{L}_{\text{primary}} + \lambda \mathcal{L}_{\text{aux}}.$$

For GoEmotions, the primary task uses sigmoid-based binary cross-entropy because the dataset is multi-label, while the auxiliary task remains single-label. If a GoEmotions instance maps to conflicting auxiliary categories, we mask that instance from the auxiliary loss while retaining it for the primary loss. Under this procedure only approximately 5–8% of training instances are masked from the auxiliary loss; exact task-level coverage is reported in Appendix A.

### 3.3 Auxiliary Supervision Sources

We compare three sources of auxiliary supervision:

- **Aligned**: auxiliary labels are derived deterministically from the original emotion labels using dataset-specific mappings.
- **Weak**: auxiliary labels are generated with VADER.
- **LLM**: auxiliary labels are generated with Qwen2.5-14B using zero-shot prompting.

In the **Aligned** setting, auxiliary labels are directly mapped from the gold emotion annotations. This setting has the highest internal consistency because the auxiliary labels are derived from the same annotations used for the primary task.

| |
|---|
| **Coarse Affective Valence** |
| *positive*: admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief |
| *negative*: anger, annoyance, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, remorse, sadness |
| *ambiguous*: confusion, curiosity, neutral, realization, surprise |
| **Valence-Activation Structure** |
| *positive*: admiration, amusement, approval, caring, curiosity, excitement, gratitude, joy, love, optimism, pride, relief |
| *negative_active*: anger, annoyance, disapproval, disgust, nervousness |
| *negative_passive*: disappointment, embarrassment, fear, grief, remorse, sadness |
| *ambiguous*: confusion, desire, neutral, realization, surprise |
| **Positive affect Detection** |
| *positive*: admiration, amusement, approval, caring, desire, excitement, gratitude, joy, love, optimism, pride, relief |
| *non_positive*: anger, annoyance, confusion, curiosity, disappointment, disapproval, disgust, embarrassment, fear, grief, nervousness, neutral, realization, remorse, sadness, surprise |

Table 1: GoEmotions auxiliary-label mappings. GoEmotions contains 27 emotion labels plus *neutral*. Dataset-specific mappings for MELD and DAIR-AI Emotion are provided in Appendix A.9.

In the **Weak** setting, VADER compound scores are converted into coarse valence labels using fixed thresholds: scores greater than 0.05 are mapped to *positive*, scores below -0.05 are mapped to *negative*, and scores in between are mapped to *ambiguous*. For Task C, negative instances are further divided into *negative_active* and *negative_passive* using the rule-based activation heuristic described in Appendix A.

In the **LLM** setting, Qwen2.5-14B is prompted to assign exactly one label from the predefined auxiliary taxonomy. We use low-temperature generation with a fixed seed, setting `temperature`=0.1 and `top-p`=0.9. Because these settings are not strictly deterministic sampling-free decoding, we refer to this setup as low-temperature generation rather than deterministic decoding. Prompt templates are provided in Appendix A.

### 3.4 Auxiliary Label Mappings

The auxiliary mappings are dataset-specific because MELD, GoEmotions, and DAIR-AI Emotion use different label inventories. Table 1 summarizes the GoEmotions mapping because it has the largest label space. Full mappings for all datasets are provided in Appendix A.9.

## 4 Experimental Setup

### 4.1 Datasets

Our main experiments evaluate three English emotion classification benchmarks:

- **MELD** (Poria et al., 2019): a conversational emotion recognition dataset derived from the TV show *Friends*, annotated with seven emotion categories. Although MELD is multimodal in its original form, we use only the textual utterances in order to keep the comparison controlled across datasets.
- **GoEmotions** (Demszky et al., 2020): a Reddit-based dataset containing approximately 58k comments annotated with 27 emotion categories plus *neutral*. We treat GoEmotions as a multi-label classification task.
- **DAIR-AI Emotion** (Saravia et al., 2018): a Twitter-based dataset containing short texts annotated with six basic emotion categories.

The datasets differ in label granularity, domain, and ambiguity, allowing us to evaluate auxiliary supervision across conversational, social-media, and multi-label settings.

As an additional robustness benchmark, we evaluate SemEval-2018 Task 1 (English, E-c) (Mohammad et al., 2018). We use the same preprocessing as for GoEmotions, with a multi-label primary task and a single-label auxiliary target, five-seed protocol and training fractions of 5%, 10%, 25%, 50%, and 100%. We evaluate aligned and VADER-weak supervision.

### 4.2 Training Procedure

The main experiments use `roberta-base` as the shared encoder. Each model consists of the shared encoder, a primary emotion-classification head, and one auxiliary head. We additionally conduct a focused robustness check with `deberta-v3-base` on MELD under the 10% training-data setting.

Unless otherwise stated, models are trained with AdamW using a learning rate of $2 \times 10^{-5}$, weight decay of 0.01, batch size 16, maximum sequence length 128, and linear warmup over the first 10% of training steps. We train for at most 10 epochs and apply early stopping with patience 3 based on validation Macro-F1. Gradients are clipped at 1.0. The auxiliary loss weight is set to $\lambda = 0.2$ in the main experiments.

For GoEmotions, the primary objective uses `BCEWithLogitsLoss` with class-specific positive weights derived from the training distribution. Predictions are obtained with a fixed threshold of 0.5. The threshold is intentionally not tuned so that all supervision sources are evaluated under the same decoding conditions. Auxiliary objectives use weighted cross-entropy.

All main-grid RoBERTa-base experiments on MELD, GoEmotions, and DAIR-AI Emotion (both full-data and low-resource) use ten random seeds: 11, 22, 33, 44, 99, 55, 66, 77, 88, and 23. The means and standard deviations reported in the main tables are computed over all ten seeds, and the paired significance tests reported in Section 5.1 also use all ten seeds. The following auxiliary experiments use five seeds (11, 22, 33, 44, 99) only: the DeBERTa-v3-base backbone-robustness runs, the MELD confusion-matrix analysis, the Random Shuffle and regularization control baselines, and the pseudo-labeling and uncertainty-weighting comparison. SemEval-2018 experiments also use five seeds (11, 22, 33, 44, 99). We report Macro-F1 as the primary evaluation metric. For MELD and DAIR-AI Emotion, Macro-F1 is computed over single-label emotion classes. For GoEmotions, Macro-F1 is computed label-wise over all 28 labels.

We compare each MTL configuration with its seed-matched single-task baseline using paired two-sided $t$-tests and assess the agreement–transfer association using Pearson's $r$ with a two-sided $p$-value. Because the $p$-values are uncorrected for multiple comparisons, we interpret them as exploratory and emphasize effect sizes, directionality, and consistency across settings.

### 4.3 Scaling Study

To evaluate low-resource behavior, we train models using 5%, 10%, 25%, 50%, and 100% of the original training data while keeping validation and test splits fixed. For single-label datasets, subsets are sampled with class-stratified sampling. For GoEmotions, which is multi-label, we use iterative stratification where available; otherwise, we preserve marginal label frequencies as closely as possible and report the resulting auxiliary-label coverage in Appendix A.

## 5 Results

### 5.1 Low-Resource Performance

Auxiliary supervision provides its clearest benefits when labeled training data are limited. Figure 1 shows the scaling trends across training-data fractions. MTL variants generally outperform the single-task baseline in low-resource regimes, especially on MELD and GoEmotions.

Representative low-resource results are shown in Table 2. On GoEmotions with 5% of the training data, the single-task baseline achieves 0.3763 Macro-F1, while strong MTL configurations reach 0.3815 with $A + D_{\text{aligned}}$ and 0.3814 with $A + D_{\text{LLM}}$. On MELD with 10% of the training data, LLM-based Task D improves Macro-F1 from 0.3914 to 0.3994.

In contrast, DAIR-AI Emotion shows smaller and less consistent improvements, with the single-task baseline remaining competitive across most training regimes. As training data increases, the gap between MTL and the single-task baseline generally narrows. These results suggest that auxiliary supervision is most useful when the primary task benefits from additional affective structure under limited supervision.

Using all ten seeds, on GoEmotions with 5% of the training data, the improvements over the single-task baseline yield nominal $p < .05$ in exploratory paired two-sided $t$-tests for $A + D_{\text{aligned}}$ ($p = 0.0069$) and $A + D_{\text{LLM}}$ ($p = 0.0341$), uncorrected for multiple comparisons.

### 5.2 Auxiliary Loss Weight Sensitivity

To assess sensitivity to auxiliary supervision strength, we evaluate $\lambda \in \{0.05, 0.1, 0.2, 0.5, 1.0\}$ on MELD with 10% of the training data using LLM-based Task D supervision and five random seeds (11, 22, 33, 44, and 99).

For increasing values of $\lambda$, Macro-F1 (mean $\pm$ standard deviation) is $0.3932 \pm 0.0191$, $0.3910 \pm 0.0115$, $0.4052 \pm 0.0165$, $0.4055 \pm 0.0106$, and $0.4018 \pm 0.0062$, respectively. Performance varies by at most $0.0145$ Macro-F1, with nearly identical results for $\lambda = 0.2$ and $\lambda = 0.5$, indicating limited sensitivity within the evaluated range.

### 5.3 Backbone Robustness

To test whether the low-resource trends are specific to RoBERTa-base, we repeat the MELD 10% experiment using DeBERTa-v3-base. The single-task DeBERTa baseline reaches $0.3818 \pm 0.0123$

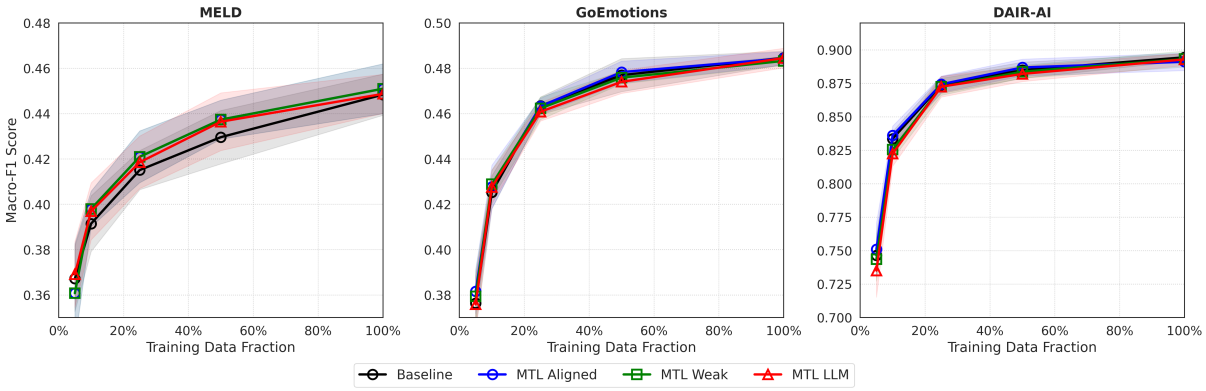



Figure 1: Scaling curves for MELD, GoEmotions, and DAIR-AI Emotion. At each training fraction, the curve for each supervision source reports the auxiliary task among B, C, D with the highest validation Macro-F1. The single-task model (Task A only) is included as the baseline. Curves show mean Macro-F1 over ten random seeds, and shaded regions indicate standard deviation. Colors, marker shapes, and line styles distinguish supervision sources.

| Dataset / Backbone | Setting | Baseline | Representative MTL | Δ Macro-F1 |
|---|---|---|---|---|
| MELD / RoBERTa-base | 10% train | 0.3914 | 0.3994 ($A + D_{\text{LLM}}$) | +0.0080 |
| GoEmotions / RoBERTa-base | 5% train | 0.3763 | 0.3815 ($A + D_{\text{aligned}}$) | +0.0052 |
| GoEmotions / RoBERTa-base | 5% train | 0.3763 | 0.3814 ($A + D_{\text{LLM}}$) | +0.0051 |
| MELD / DeBERTa-v3-base | 10% train | 0.3818 | 0.3922 ($A + C_{\text{LLM}}$) | +0.0104 |
| MELD / DeBERTa-v3-base | 10% train | 0.3818 | 0.3916 ($A + B_{\text{LLM}}$) | +0.0098 |

Table 2: Representative low-resource and backbone-robustness results. Baseline refers to single-task emotion classification ($A$ only). Representative MTL denotes a high-performing auxiliary-supervision configuration in the corresponding setting. RoBERTa-base rows are averaged over ten random seeds; DeBERTa-v3-base rows are averaged over five random seeds and are included as a focused robustness check rather than a full backbone sweep.

Macro-F1. The strongest auxiliary configurations again improve over the baseline: LLM-based Task C reaches $0.3922 \pm 0.0075$, and LLM-based Task B reaches $0.3916 \pm 0.0048$.

These gains are modest and are not statistically significant across five seeds, but they follow the same direction as the RoBERTa-base results. This suggests that the observed low-resource trend is not tied to one encoder family, although a broader backbone sweep would be needed for stronger generalization claims.

### 5.4 Additional Benchmark: SemEval-2018 Task 1

On SemEval-2018 Task 1, results follow the same overall trend as in the main experiments, with small improvements at 10% ($0.5482 \rightarrow 0.5513$ with $A + D_{\text{aligned}}$) and 50% ($0.5829 \rightarrow 0.5854$ with $A + D_{\text{weak}}$). At the remaining training fractions the single-task baseline is comparable to the best MTL configuration: at 5% ($0.5116 \rightarrow 0.5118$ with $A + D_{\text{aligned}}$), at 25% ($0.5748 \rightarrow 0.5759$ with $A + D_{\text{weak}}$), and at 100% ($0.5950 \rightarrow 0.5953$ with $A + B_{\text{aligned}}$). Consistent with the other datasets, gains from auxiliary supervision on SemEval-2018 are marginal and do not exceed roughly 0.3 points Macro-F1.

### 5.5 Error Reduction Analysis

To understand how auxiliary supervision changes prediction behavior, we analyze confusion-matrix differences between the baseline and MTL systems:

$$\Delta CM = CM_{\text{MTL}} - CM_{\text{Base}}.$$

For off-diagonal cells, negative values indicate reduced confusion relative to the baseline. For diagonal cells, negative values indicate fewer correct predictions and should therefore be interpreted differently. Confusion-matrix deltas are computed from predictions aggregated across the five evaluation seeds. Representative MELD confusion-matrix heatmaps are provided in Appendix A.11.

On MELD, auxiliary supervision reduces several confusions among semantically related emotion categories. In particular, Task B with aligned or weak supervision reduces confusion between *disgust* and *anger*, and between *fear* and *sadness*. LLM-based

Task D is also associated with reduced *joy* → *neutral* confusion, suggesting improved recognition of subtle positive expressions.

Overall, the error analysis indicates that auxiliary supervision is most helpful for categories with semantic overlap or high baseline confusion.

### 5.6 Per-Class Analysis

To examine whether auxiliary supervision affects all emotion categories uniformly, we analyze per-class F1 changes on GoEmotions under the 10% training-data setting. This analysis complements the aggregate Macro-F1 results by showing which categories account for the observed improvements.

The largest changes are concentrated among low-frequency or semantically overlapping categories, including *remorse* and *embarrassment*. These categories are difficult under the single-task baseline and benefit more from auxiliary affective structure than high-frequency or semantically broader categories. In contrast, frequent categories tend to show smaller relative changes, suggesting that auxiliary supervision is most useful where the primary task provides limited or ambiguous evidence.

We also observe that LLM-based auxiliary supervision produces higher variance for some low-frequency categories, including sizeable degradations for classes such as *pride*, *grief*, and *fear*, while categories such as *embarrassment* improve under aligned supervision but degrade under LLM supervision. This indicates that noisy auxiliary labels may provide useful affective structure in aggregate while remaining less stable for rare classes. Overall, the per-class results support the broader finding that the largest effects of auxiliary supervision—both improvements and degradations—occur among low-frequency or semantically overlapping categories.

Full per-class F1 changes are provided in Appendix A.10.

### 5.7 Additional Baselines and Robustness Checks

We further compare against control baselines designed to test whether the observed MTL gains are due to semantically meaningful auxiliary labels or to more generic regularization effects. We evaluate three controls on MELD under the 10% training-data setting: increased dropout, label smoothing, and a Random Shuffle auxiliary baseline.

The Random Shuffle baseline preserves the auxiliary taxonomy and marginal label frequencies while removing the semantic correspondence between utterances and auxiliary labels; implementation details are provided in Appendix A.12.

In this five-seed control comparison on MELD 10%, the single-task baseline achieves 0.3944, increased dropout 0.3368, label smoothing 0.3526, Random Shuffle $0.3707 \pm 0.0130$, aligned Task D 0.3981, and LLM-based Task D 0.4052. All three control baselines fall below both the single-task baseline and semantically grounded auxiliary supervision. These specific controls therefore do not reproduce the observed gains, although they do not rule out all regularization- or optimization-based explanations.

These comparisons do not fully isolate all optimization effects associated with multi-task learning, but they indicate that the observed improvements are not reproduced by adding an auxiliary objective with the same marginal label distribution or by generic regularization alone. Semantically structured auxiliary supervision appears to contribute more effectively to downstream transfer in low-resource settings.

In a separate five-seed comparison, we situate our framework relative to alternative low-resource strategies by evaluating pseudo-labeling and uncertainty-based multi-task weighting on the representative MELD 10% setting using seeds 11, 22, 33, 44, 99. Pseudo-labeling reaches $0.411$ Macro-F1, uncertainty weighting reaches $0.405$, and our best MTL model reaches $0.405$ under this same five-seed protocol (which is why the MTL value is $0.405$ here rather than the ten-seed value of $0.3994$ reported in Table 2). These results are comparable, while our framework remains lightweight and simple to implement.

To assess whether the observed MTL gains are robust to noise in the auxiliary supervision, we perturb the LLM-based Task D auxiliary labels on MELD 10% at controlled rates $\{5\%, 10\%, 20\%\}$ and retrain under the main protocol (ten seeds, ten epochs, patience 3, $\lambda = 0.2$; a fresh perturbation realization per seed). Under this matched protocol, MTL Macro-F1 is $0.3994 \pm 0.0127$ with unperturbed LLM auxiliary labels, $0.3963 \pm 0.0152$ at 5% perturbation, $0.3938 \pm 0.0079$ at 10%, and $0.3919 \pm 0.0097$ at 20%, compared to the single-task baseline of $0.3914 \pm 0.0123$ from Table 2 (Table 3). Mean MTL Macro-F1 decreases monotonically with the perturbation rate, but remains at or above the single-task baseline at every rate tested, including at 20% perturbation (paired-$t$ vs.

| Perturbation | MTL Macro-F1 | Δ vs. STL |
|---|---|---|
| 0% (unperturbed) | $0.3994 \pm 0.0127$ | $+0.0080$ |
| 5% | $0.3963 \pm 0.0152$ | $+0.0048$ |
| 10% | $0.3938 \pm 0.0079$ | $+0.0024$ |
| 20% | $0.3919 \pm 0.0097$ | $+0.0005$ |
| STL baseline | $0.3914 \pm 0.0123$ | — |

Table 3: Sensitivity of MTL to auxiliary-label perturbation on MELD 10% with LLM-based Task D supervision. Reported as mean ± standard deviation over ten seeds under the main training protocol. STL baseline is taken from Table 2. Mean MTL Macro-F1 decreases monotonically with the perturbation rate but remains numerically at or above the single-task baseline throughout, though the margin becomes small and is not statistically significant.

STL: $p = 0.107$, $0.402$, $0.429$, $0.920$ for rates $0.00/0.05/0.10/0.20$).

### 5.8 Representation and Gradient Analysis

To probe the transfer mechanism, we perform representation-geometry and gradient-similarity analyses on MELD, comparing the shared encoder under the single-task baseline and under MTL. The representation analysis shows that MTL increases the separation between historically confused emotion classes by up to 19.02% for *fear* vs. *sadness* (baseline centroid distance $5.499$ vs. MTL $6.545$) and 15.17% for *disgust* vs. *anger* (baseline $6.543$ vs. MTL $7.535$). The gradient-similarity analysis measures the mean cosine similarity between primary-task and auxiliary-task gradients on the shared encoder for Task B under the three supervision sources. The mean cosine similarities are small in magnitude and vary in sign across sources: $-0.056 \pm 0.095$ for aligned supervision, $-0.045 \pm 0.157$ for VADER-weak supervision, and $+0.008 \pm 0.109$ for LLM supervision. The gradients are therefore close to orthogonal on average and do not systematically conflict, indicating that the auxiliary objective contributes an approximately independent training signal rather than reinforcing or opposing the primary gradient direction.

## 6 Discussion

### 6.1 Label Agreement and Transfer Utility

A central question in this study is whether higher auxiliary-label agreement necessarily leads to better downstream transfer. To examine this relationship, we compute agreement between each auxiliary supervision source and the aligned gold-derived labels using Cohen's $\kappa$, then compare agreement with downstream improvement in Macro-F1.

As shown in Figure 2, we find no clear statistically supported association between agreement and transfer utility: the Pearson correlation is $r = -0.30$ with a two-sided $p = 0.134$ (uncorrected), computed on 10-seed full-data deltas. Although higher-agreement supervision sources sometimes perform well, this analysis does not establish agreement as a reliable predictor of downstream improvement.

The clearest counterexample appears in MELD under low-resource conditions: LLM-generated auxiliary labels show relatively low agreement with aligned labels ($\kappa \approx 0.21$–$0.26$) while still producing competitive or improved downstream performance. This suggests that auxiliary supervision may help not only by reproducing gold-derived mappings, but also by encouraging the model to encode broader affective structure.

Together with the error analysis, the representation and gradient analyses are consistent with the hypothesis that even low-agreement supervision can improve the shared representation by providing useful structural signal during representation learning rather than by reproducing gold-derived mappings.

### 6.2 When Does Auxiliary Supervision Help

Across datasets, auxiliary supervision is most helpful in low-resource settings and for semantically overlapping categories, consistent with the view that auxiliary objectives act as inductive biases that encourage the shared encoder to organize inputs according to coarse affective structure when the primary task does not provide enough labeled examples. The gains are not uniform: DAIR-AI Emotion shows smaller and less consistent improvements, suggesting that auxiliary supervision is most useful when the primary label space is ambiguous, imbalanced, or fine-grained, as in MELD and GoEmotions.

### 6.3 Implications for LLM-Generated Supervision

Our results suggest a cautious but useful role for LLM-generated labels: they need not be reliable enough to replace human annotations in order to provide useful auxiliary structure when optimized jointly with gold primary labels. This should not be read as evidence that synthetic labels are

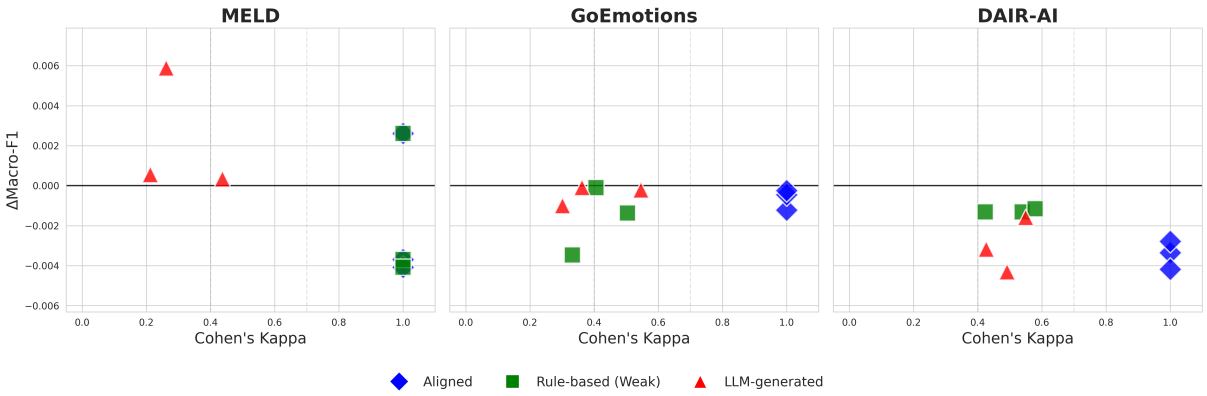



Figure 2: Relationship between auxiliary-label agreement, measured by Cohen's $\kappa$ against aligned labels, and downstream improvement in Macro-F1. Each point corresponds to one (dataset, auxiliary task $\in \{B, C, D\}$, supervision source) combination on the full-data setting; colors and marker styles indicate the supervision source. In this analysis we find no clear statistically supported association between agreement and transfer utility.

universally beneficial—LLM supervision remains noisy, can vary across rare classes, and may encode systematic biases. The appropriate conclusion is therefore narrow: LLM-generated auxiliary labels can be useful in some low-resource emotion-classification settings, but their utility is dataset-dependent and should be evaluated empirically.

## 7 Conclusion

We presented a controlled study of auxiliary supervision for low-resource emotion classification. Using a shared-encoder MTL framework, we compared three auxiliary objectives and three supervision sources across MELD, GoEmotions, and DAIR-AI Emotion.

The results show that auxiliary supervision provides its clearest benefits in low-resource settings, particularly on MELD and GoEmotions. Error and per-class analyses further show that the largest effects—both improvements and degradations—occur among low-frequency or semantically overlapping categories. We also find no clear statistically supported association between auxiliary-label agreement and downstream transfer, indicating that noisy auxiliary labels can still be useful when they preserve relevant affective structure.

Overall, our findings support auxiliary supervision as a lightweight and practical inductive bias for low-resource emotion classification. At the same time, the gains are modest and dataset-dependent, underscoring the need for careful auxiliary-task design, transparent reporting, and evaluation beyond aggregate performance.

## 8 Ethical Considerations

This work uses public English-language datasets from social media and dialogue sources, which may contain demographic, cultural, and annotation biases. Emotion labels are inherently subjective and should not be interpreted as reliable indicators of a speaker's internal mental state.

LLM-generated auxiliary labels may additionally reflect biases of the underlying model or oversimplify culturally specific emotional expressions. In this work, such labels are used only as auxiliary supervision rather than replacements for gold labels.

## 9 Limitations

This study has several limitations. First, all experiments are conducted on English-language datasets, so the conclusions may not generalize to other languages, domains, or cultural settings. Emotion expression is culturally variable, and the auxiliary taxonomies used here may not transfer directly across languages.

Second, we use text-only inputs for all datasets, including MELD. This design keeps the comparison controlled across datasets, but it means that our results should not be interpreted as multimodal conversational emotion recognition results.

Third, the auxiliary objectives are intentionally simple abstractions based on valence and activation. More complex auxiliary objectives, including continuous affect representations, learned task groupings, or dynamically weighted auxiliary losses, may produce different results.

Fourth, the empirical effects are modest and

sometimes inconsistent. Although the main experiments use ten random seeds and multiple training fractions, the number of seeds is still small relative to the number of comparisons. Statistical tests should therefore be interpreted as exploratory rather than confirmatory.

Fifth, we evaluate one main LLM family for synthetic auxiliary labeling. Other LLMs, decoding settings, prompt formats, or calibration strategies may yield different auxiliary-label distributions and transfer behavior.

Finally, while the Random Shuffle, gradient-similarity, and representation-geometry analyses suggest that semantically meaningful auxiliary structure matters, they do not establish a full causal mechanism for why certain auxiliary labels help. Future work should examine gradient interactions, representation structure, and prompt sensitivity more directly.

## Acknowledgements

This work was supported by the Slovak Research and Development Agency under the Contract no. APVV-22-0414.

## References

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. 2020. Pseudo-labeling and confirmation bias in deep semi-supervised learning. *Preprint*, arXiv:1908.02983.

Jeremy Barnes. 2023. Sentiment and emotion classification in low-resource settings. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 290–304, Toronto, Canada. Association for Computational Linguistics.

Thomas Borsani, Andrea Rosani, Giuseppe Nicosia, and Giuseppe Di Fatta. 2026. Gradient similarity surgery in multi-task deep learning. In *Machine Learning and Knowledge Discovery in Databases. Research Track and Applied Data Science Track*, pages 95–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Shijie Chen, Yu Zhang, and Qiang Yang. 2024. Multi-task learning in natural language processing: An overview. *ACM Comput. Surv.*, 56(12).

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 794–803. PMLR.

Lukas Christ and Shahin Amiriparian. 2026. Training-free text emotion tagging via LLM-based best-worst scaling. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 6678–6694, Rabat, Morocco. Association for Computational Linguistics.

Simon Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *Preprint*, arXiv:2009.09796.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6748–6761, Abu Dhabi, UAE. Association for Computational Linguistics.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: A contrast-enhanced prompt-tuning framework for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2947–2957, Torino, Italia. ELRA and ICCL.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020.

Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Junqing Gong, Binhan Yang, and Wei Shen. 2025. A triple-view framework for fine-grained emotion classification with clustering-guided contrastive learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4970–4984, Vienna, Austria. Association for Computational Linguistics.

Lynn Greschner and Roman Klinger. 2025. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and LLMs. In *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, pages 628–646, Albuquerque, USA. Association for Computational Linguistics.

C. J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 216–225.

Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2026. Linguistic signatures for enhanced emotion detection. In *Proceedings of the ACM Web Conference 2026*, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *Preprint*, arXiv:1907.11692.

Navonil Majumder, Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. 2024. From text to emotion: Unveiling the emotion annotation capabilities of llms. In *Interspeech 2024*, pages 2650–2654.

Nicholas Pangakis and Sam Wolken. 2024. Knowledge distillation in automated annotation: Supervised text classification with LLM-generated training labels. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 113–131, Mexico City, Mexico. Association for Computational Linguistics.

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.

Xiaohan Qin, Xiaoxing Wang, and Junchi Yan. 2025. Towards consistent multi-task learning: Unlocking the potential of task-specific parameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10067–10076.

Md Mezbaur Rahman and Cornelia Caragea. 2025. LLM-guided co-training for text classification. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31104–31121, Suzhou, China. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2023. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153.

Trevor Standley, Amir R. Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. 2020. Which tasks should be learned together in multi-task learning? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9120–9132. PMLR.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and

synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Juntao Wang and Tsunenori Mine. 2023. Multi-task learning for emotion recognition in conversation with emotion shift. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 257–266, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. 2020. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 5824–5836.

# A Reproducibility Details

## A.1 Training Hyperparameters

Table 4 lists the key hyperparameters used in the main experiments.

| Hyperparameter | Value |
|---|---|
| Backbone | `roberta-base` |
| Robustness backbone | `deberta-v3-base` (MELD 10%) |
| Optimizer | AdamW |
| Learning rate | $2 \times 10^{-5}$ |
| Weight decay | 0.01 |
| Warmup fraction | 0.1 |
| Max sequence length | 128 |
| Batch size | 16 |
| Max epochs | 10 |
| Early stopping patience | 3 |
| Gradient clipping | 1.0 |
| Auxiliary loss weight | $\lambda = 0.2$ |
| GoEmotions threshold | 0.5 |
| Evaluation seeds (main) | 11, 22, 33, 44, 99, 55, 66, 77, 88, 23 |
| SemEval-2018 seeds | 11, 22, 33, 44, 99 |

Table 4: Training hyperparameters used across experiments.

## A.2 Computational Infrastructure

The main experiments use `roberta-base`, which contains approximately 125M parameters. The focused robustness check uses `deberta-v3-base`, which contains approximately 184M parameters. Experiments were run on the HPC cluster using NVIDIA H200 GPUs. We estimate the total compute budget to be approximately 120 GPU-hours across model training, low-resource scaling runs, auxiliary-loss tuning, control baselines, and the DeBERTa-v3-base robustness check.

Auxiliary-label generation with Qwen2.5-14B was performed once before model training, and the resulting labels were reused across training runs.

## A.3 LLM Prompt Templates

Synthetic auxiliary labels are generated using zero-shot prompting with Qwen2.5-14B. The model is instructed to output exactly one label from the specified taxonomy.

## A.4 Task B (Coarse Affective Valence)

Classify the coarse affective valence of the following text into exactly one category: {positive, negative, ambiguous}

Text: {text}

Label:

## A.5 Task C (Valence–Activation Structure)

Classify the affective valence and activation level of the following text into exactly one category: {positive, negative_active, negative_passive, ambiguous}

Text: {text}

Label:

## A.6 Task D (Positive Affect Detection)

Classify whether the following text expresses positive affect into exactly one category: {positive, non_positive}

Text: {text}

Label:

## A.7 Weak Supervision Details

For VADER-based weak supervision, we first compute the VADER compound sentiment score for each input text. Following the standard VADER thresholds, texts with compound scores greater than 0.05 are assigned to *positive*, texts with compound scores below -0.05 are assigned to *negative*, and texts with scores in the interval $[-0.05, 0.05]$ are assigned to *ambiguous*.

For Task B, these three labels are used directly as the auxiliary targets.

For Task D, the VADER-derived labels are collapsed into a binary positive-affect target: *positive*

is mapped to *positive*, while both *negative* and *ambiguous* are mapped to *non_positive*.

For Task C, we further split VADER-derived negative instances into *negative_active* and *negative_passive*. This split is based on a deterministic lexical activation heuristic. Negative texts are assigned to *negative_active* if they contain at least one token from an anger-, disgust-, or agitation-related keyword list, or if they contain strong surface markers of activation such as repeated exclamation marks or all-caps emphasis. Otherwise, they are assigned to *negative_passive*. Non-negative texts retain their Task B labels, with *positive* mapped to *positive* and *ambiguous* mapped to *ambiguous*.

The activation keyword list is:

> angry, anger, annoyed, annoying, annoy, furious, fury, mad, hate, hated, hateful, disgust, disgusting, disgusted, gross, awful, terrible, horrible, pissed, rage, raging, irritated, irritating, upset, outraged, offensive, attack, attacked, attacking, blame, stupid, idiot, damn

Before keyword matching, text is lowercased and tokenized using whitespace and punctuation splitting. Keyword matching is performed on full tokens rather than substrings. All heuristic labels are generated before model training and remain fixed across random seeds. This procedure is intentionally simple: the weak labels are not intended to approximate gold auxiliary labels perfectly, but to provide a reproducible rule-based source of coarse affective supervision.

### A.8 GoEmotions Auxiliary-Label Masking

Because GoEmotions is multi-label, some examples map to conflicting auxiliary labels. These examples are masked from the auxiliary loss using `ignore_index=-100` while remaining available for the primary emotion-classification objective. The masking procedure preserves 92.97% of training instances for Task B, 91.62% for Task C, and 95.00% for Task D. The resulting auxiliary-label distributions remain highly correlated with the original distribution, with Pearson $r > 0.99$.

### A.9 Auxiliary Label Mappings

This section reports the dataset-specific mappings used to derive aligned auxiliary labels for MELD (Table 5) and DAIR-AI Emotion (Table 6). The GoEmotions mapping is shown in Table 1.

| Emotion | Task B | Task C / Task D |
|---|---|---|
| joy | positive | positive / positive |
| surprise | ambiguous | ambiguous / non_positive |
| neutral | ambiguous | ambiguous / non_positive |
| anger | negative | negative_active / non_positive |
| disgust | negative | negative_active / non_positive |
| fear | negative | negative_passive / non_positive |
| sadness | negative | negative_passive / non_positive |

Table 5: Aligned auxiliary-label mapping for MELD.

| Emotion | Task B | Task C / Task D |
|---|---|---|
| joy | positive | positive / positive |
| love | positive | positive / positive |
| anger | negative | negative_active / non_positive |
| fear | negative | negative_passive / non_positive |
| sadness | negative | negative_passive / non_positive |
| surprise | ambiguous | ambiguous / non_positive |

Table 6: Aligned auxiliary-label mapping for DAIR-AI Emotion.

### A.10 Additional Per-Class Results

Figure 3 reports per-class F1 changes for GoEmotions under the 10% training-data setting. The analysis complements the aggregate Macro-F1 results by showing that auxiliary supervision affects emotion categories unevenly: the largest changes—both improvements and degradations—are concentrated among low-frequency or semantically overlapping classes such as *pride*, *grief*, *fear*, and *embarrassment*.

### A.11 MELD Confusion-Matrix Heatmaps

Figure 4 shows the baseline MELD confusion matrix and representative confusion-matrix deltas for auxiliary-supervision models. The baseline matrix summarizes prediction errors of the single-task emotion classifier, while the delta heatmaps show changes after adding auxiliary supervision. For off-diagonal cells, negative values indicate reduced confusion relative to the single-task baseline. For diagonal cells, positive values indicate more correct predictions and negative values indicate fewer correct predictions.

### A.12 Random Shuffle and Regularization Baselines

For the Random Shuffle baseline, Task D auxiliary labels on MELD were randomly permuted independently for each evaluation seed while preserving the original class distribution. This procedure removes the semantic correspondence between utterances and auxiliary labels while maintaining identical

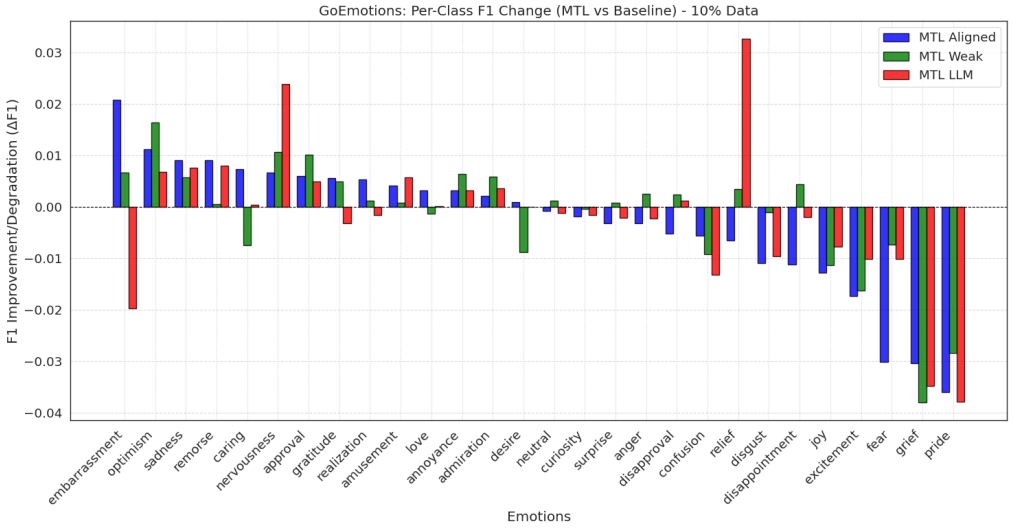
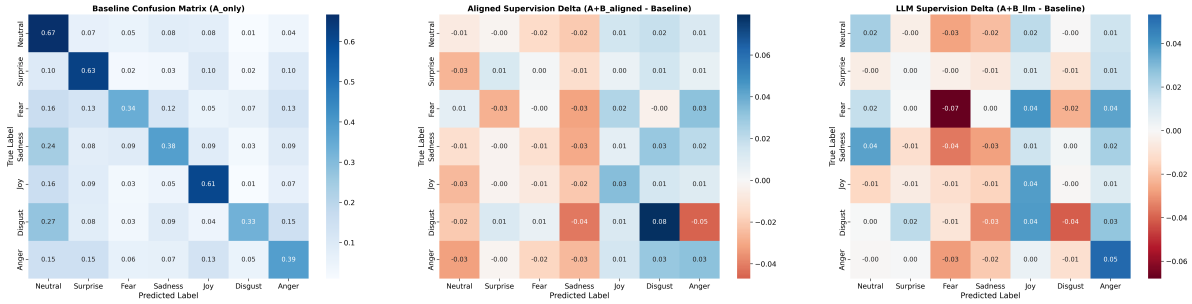



Figure 3: Per-class F1 changes for GoEmotions under the 10% training-data setting relative to the single-task baseline, aggregated over the best auxiliary task per source. Positive values indicate improved per-class F1 under auxiliary supervision and negative values indicate degradations. Colors correspond to the auxiliary supervision sources used throughout the paper. Effects are heterogeneous across classes: several low-frequency or semantically overlapping categories (e.g., *pride*, *grief*, *fear*, *embarrassment*) exhibit large changes in both directions depending on the supervision source.




(a) Baseline (b) Aligned Task B $\Delta CM$ (c) LLM Task D $\Delta CM$

Figure 4: Confusion-matrix analysis for MELD. The baseline matrix shows common confusions among emotion categories. Delta heatmaps show changes relative to the single-task baseline, where negative off-diagonal values indicate reduced confusion under auxiliary supervision.

auxiliary-label frequencies. The baseline therefore tests whether improvements arise from meaningful auxiliary structure rather than from adding an auxiliary loss with the same marginal label distribution.

We also evaluated two regularization baselines, increased dropout and label smoothing, to compare auxiliary supervision against generic regularization. These baselines used the same optimizer configuration as the primary experiments, including AdamW optimization and a learning rate of $2 \times 10^{-5}$. The increased-dropout baseline used dropout probability 0.3, and the label-smoothing baseline used smoothing coefficient 0.1.