STRUCTURED AUXILIARY SUPERVISION FOR CONTINUOUS IMBALANCED FINANCIAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Financial sentiment regression with large language models (LLMs) trains on neutral-dominated data in which the extreme scores that drive financial decisions are rare, so fine-tuned models pull their predictions toward the dense center, a failure mode we term neutral collapse. Existing remedies reweight samples by label density or classify over a partition fixed before training, leaving the category structure of the continuous label space unlearned, and an auxiliary classification task added to supply it interferes with regression and converges at a different rate, so no fixed task weight serves both. We propose Structured Auxiliary Supervision (SAS), a plug-and-play framework that makes the label partition a trained parameter and schedules its supervision by the convergence trajectory of each task. Differentiable Structured Partitioning (DSP) learns thresholds that split the sentiment space into balanced auxiliary categories through differentiable soft assignments, giving the sparse extremes their own supervision targets. Trajectory-based Convergence Balancing (TCB) sets the task weights from windowed convergence velocities, shifting gradient toward the slower-converging task, and weights the auxiliary categories by a softmax over their losses. The two components are coupled in one loss and one backward pass, and the auxiliary head is discarded after training, so SAS adds no inference cost to the backbone. On three financial datasets and three backbone scales, fixed-weight multi-task fine-tuning (Fixed-MTL) is worse than Single-task regression in all nine settings, whereas SAS attains the lowest MSE in all nine, 10.5% below Fixed-MTL, a reduction that stays between 9.9% and 11.6% across a 24-fold increase in backbone size. On RoBERTa-base it has the lowest MSE among twelve baselines under identical protocols, and on public financial and cross-domain benchmarks it lowers MSE by 9.4% to 10.7%; DSP alone yields 52% of the joint MSE gain and TCB alone 30%. On NEU/RoBERTa-base, SAS compresses the extreme-to-neutral MSE ratio from 3.85× to 1.95×, and its advantage widens as imbalance grows: under 20× label skew, its MSE rises by 25% versus 80% for Single-task regression and 92% for Fixed-MTL.

1 INTRODUCTION

Sentiment analysis has become essential in the financial sector for tasks such as forecasting stock market trends, predicting cryptocurrency prices, and analyzing exchange rates (Balasubramanian et al., 2022; Nezhad et al., 2026; Shao et al., 2025). Unlike conventional sentiment classification that assigns discrete labels, financial applications increasingly demand sentiment polarity analysis over $[-1, 1]$, where the magnitude encodes sentiment intensity (Tetlock, 2007; Tetlock et al., 2008). Extreme sentiment signals often trigger immediate portfolio adjustments, moderate sentiments inform gradual position changes, while neutral readings carry minimal actionable information.

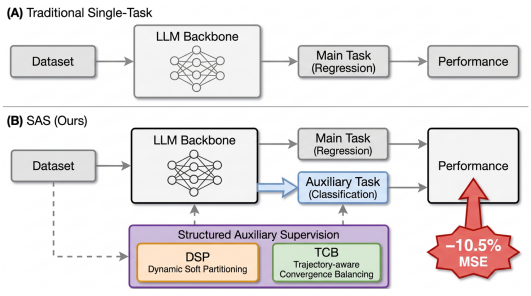


Figure 1: **Top:** standard fine-tuning regresses the score directly and collapses toward the neutral center. **Bottom:** SAS adds an auxiliary task that DSP partitions and TCB schedules; the badge is relative to Fixed-MTL.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

However, fine-tuning LLMs for financial sentiment analysis degrades under severe data distribution imbalance (He & Garcia, 2009; Branco et al., 2016; Malo et al., 2014). Financial datasets typically exhibit heavy concentration in the neutral range while extreme sentiments remain underrepresented (Omarkhan et al., 2021; Davidovic & McCleary, 2025), causing models to develop strong bias toward over-represented regions and poor prediction accuracy for critical extreme sentiment values (Mujahid et al., 2024; Nassr et al., 2025). The result is a systematic failure mode, **neutral collapse**: model predictions cluster around the neutral region, with prediction error in extreme regions up to $3.85\times$ higher than in neutral regions under Fixed-MTL.

Traditional approaches to data imbalance, such as data augmentation (Assunção et al., 2024), cost-sensitive learning (Araf et al., 2024), and threshold-moving (Leevy et al., 2023), are designed for discrete classification with well-defined category boundaries. Recent work on continuous regression imbalance either reweights the regression objective (Yang et al., 2021; Ren et al., 2022) or classifies over a partition of the target fixed before training, as in hierarchical classification adjustment (Xiong & Yao, 2024) and group classification (Pu et al., 2025); ordinal regression likewise imposes a fixed rank structure. Existing work thus treats continuous imbalance as a *reweighting* problem or as classification over a partition chosen in advance, leaving the label-space structure itself unlearned; the deficit is *structural*, and repairing it requires learning the missing category structure jointly with the model. Adding an auxiliary classification task to inject such structure, however, creates further difficulties: the regression and classification objectives interfere with each other (Yu et al., 2020), their convergence dynamics are asymmetric (Chen et al., 2018), and effective task balancing is non-trivial (Kendall et al., 2018).

To address both challenges, SAS (Figure 1) adds two coupled components to LLM fine-tuning. DSP jointly learns thresholds that partition the continuous space into balanced discrete categories, providing structural supervision for the sparse extremes. TCB sets the task weights from windowed convergence velocities, shifts gradient toward the slower-converging task, and reweights the auxiliary categories by their losses. The key insight is that the label partition can be a trained parameter: DSP learns where the categories lie and TCB decides when their supervision matters, in one loss and one backward pass, which turns continuous imbalance into a structured multi-task optimization. The main contributions are as follows:

- We identify two challenges: at the data level, distributional skew causes neutral collapse; at the task level, auxiliary task interference and convergence asymmetry prevent multi-task balancing.
- SAS, an end-to-end framework, reframes auxiliary classification as a structural lever for imbalanced regression: DSP constructs balanced discrete categories via learnable thresholds and TCB adapts task weights to windowed convergence trajectories, in one loss and one backward pass, adding no inference cost to the deployed backbone.
- SAS's advantage scales with imbalance severity: under $20\times$ skew, MSE inflates only 25% versus 80% for Single-task and 92% for Fixed-MTL, a $3.2\times$ to $3.7\times$ slower degradation.

2 MOTIVATION

Financial sentiment regression with fine-tuned LLMs faces two challenges that limit the effectiveness of current imbalance remedies.

Data-Level Challenge: Distributional Skew and Neutral Collapse. Figure 2 (left) shows that all three private financial datasets exhibit severe distributional skew, with over 60% of samples concentrated in the neutral region and fewer than 7% beyond ± 0.6. Figure 2 (middle) reveals the consequence: predictions collapse toward the neutral center, producing a distribution dramatically narrower than ground truth with severely under-covered tails. Figure 2 (right) quantifies this effect across seven sentiment regions: extreme-region MSE is $\approx 3.85\times$ higher than neutral-region MSE under Fixed-MTL, confirming that error scales monotonically with sentiment extremity. The continuous label space $[-1, 1]$ lacks discrete category boundaries, and the loss landscape is dominated by the dense neutral region. This raises a *data-level challenge*: distributional skew and neutral collapse.

Task-Level Challenge: Auxiliary Task Interference and Convergence Asymmetry. Adding a classification task alongside regression introduces two intertwined problems. Figure 3 (left): no single fixed weight w_c simultaneously optimizes both MSE and accuracy. Figure 3 (middle): the

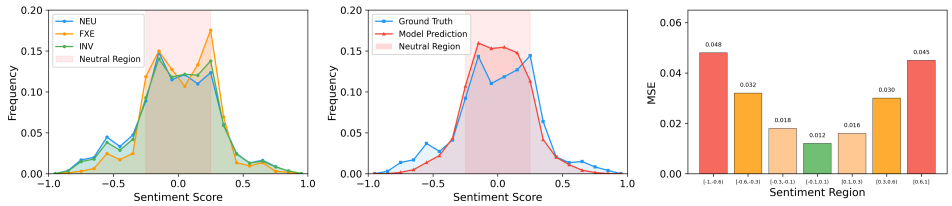


Figure 2: Neutral collapse in financial sentiment regression. **(Left)** Training distributions. **(Middle)** Prediction vs. ground truth. **(Right)** Per-region MSE.

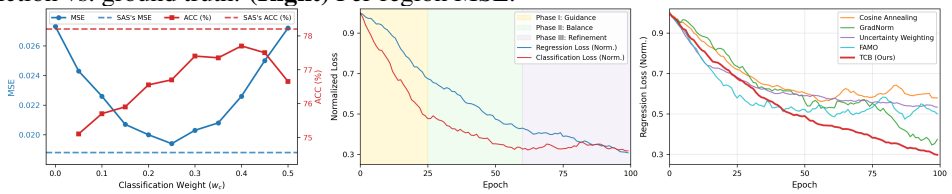


Figure 3: Auxiliary task interference and convergence asymmetry. **(Left)** Fixed-weight sensitivity. **(Middle)** Normalized loss trajectories. **(Right)** Dynamic weighting comparison.

two tasks exhibit convergence asymmetry; classification converges rapidly early while regression improves steadily throughout training. Figure 3 (right): existing dynamic weighting methods (cosine annealing (Loshchilov & Hutter, 2017), GradNorm (Chen et al., 2018), uncertainty weighting (Kendall et al., 2018), FAMO (Liu et al., 2023)) each exhibit distinct failure modes (rebounds, oscillations, plateaus). The optimal task balance shifts throughout training, and the task weights must therefore derive from the observed convergence trajectory itself. This raises a *task-level challenge*: auxiliary task interference and convergence asymmetry.

Neither challenge yields to sample-level reweighting alone: the first calls for learning the missing category structure of the label space, the second for scheduling how that structure is injected during training. These two requirements define the two components of SAS.

3 Problem Formulation

The two challenges from §2 define a joint optimization over the model parameters θ and the threshold parameters $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{K-1})$ that define the auxiliary categories, i.e., the learned label-space structure:

$$\min_{\theta, \boldsymbol{\tau}} \; \mathcal{L}(\theta, \boldsymbol{\tau}, t) = \lambda_r(t) \cdot L_r(\theta) + \lambda_c(t) \cdot L_c(\theta, \boldsymbol{\tau}) + \mathcal{R}(\boldsymbol{\tau}), \tag{1}$$

where $L_r(\theta)$ is the regression loss, $L_c(\theta, \boldsymbol{\tau})$ the auxiliary classification loss (refined in §4.3 with per-category weights q_k), $\mathcal{R}(\boldsymbol{\tau})$ a regularizer (applied independently of the task weights to prevent threshold drift), t the training step, and $\lambda_r(t), \lambda_c(t)$ adaptive task weights. The first sub-problem optimizes $\boldsymbol{\tau}$ through $L_c + \mathcal{R}$ with q_k absorbing residual within-partition difficulty (data-level challenge); the second determines $\lambda_r(t), \lambda_c(t)$ (task-level challenge).

4 Method

4.1 Structured Auxiliary Supervision Framework

To solve Eq. 1, we propose SAS, as illustrated in Figure 4. SAS attaches an auxiliary classification head and a unified training loss to an LLM backbone, and comprises two components: DSP addresses the data-level challenge by learning the label partition that defines the auxiliary targets, and TCB addresses the task-level challenge by scheduling when this supervision enters the gradient.

The components form one system because the two challenges are coupled: the partition defines the tasks whose convergence TCB balances, and the balanced gradients in turn shape the learned thresholds. A fixed partition lets the auxiliary signal drift from regression, and balancing under a misplaced partition only redistributes a misspecified signal; SAS therefore learns the partition and its schedule in one loss. In financial text this coupling matters most at the extremes: neutral-dense, extreme-sparse labels concentrate economic value where unstructured auxiliary signals under-supervise, and DSP gives these regions their own categories while TCB keeps their supervision active.

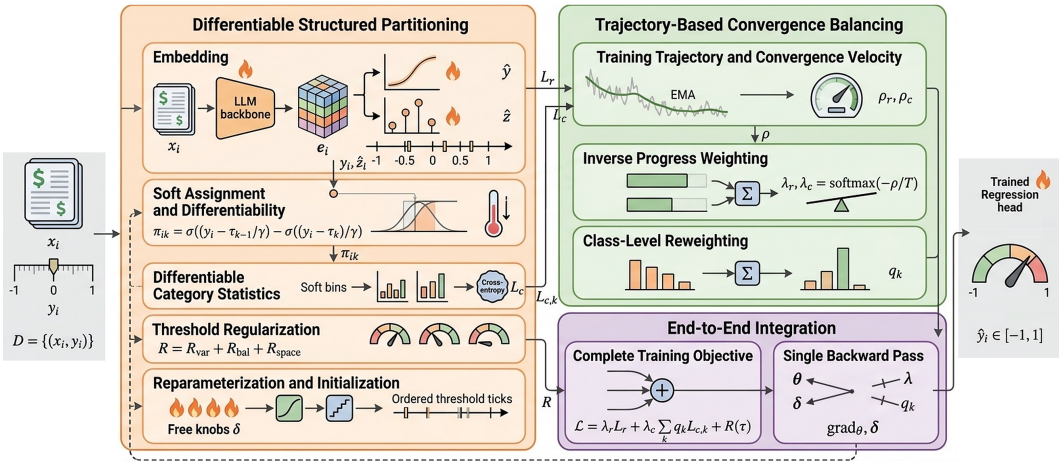


Figure 4: **Left:** training pairs (x_i, y_i), $y_i \in [-1, 1]$. **Middle:** DSP turns learned thresholds $\boldsymbol{\tau}$ into soft auxiliary targets π_{ik}. **Right:** TCB turns EMA loss trajectories into task weights λ_r, λ_c and class weights q_k. **Bottom:** a single backward pass updates $\theta, \boldsymbol{\delta}$; inference uses the regression head.

4.2 Differentiable Structured Partitioning

To address the data-level challenge, we propose DSP, as illustrated in Figure 4 (middle). It partitions the label space with learnable thresholds, turns each score into differentiable soft category targets and statistics, and regularizes the thresholds toward homogeneous, balanced categories.

Embedding. Given dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ with $y_i \in [-1, 1]$, an LLM backbone $\mathcal{M}$ maps each x_i to embedding $\mathbf{e}_i \in \mathbb{R}^d$. A regression head produces $\hat{y}_i = h_r(\mathbf{e}_i) \in [-1, 1]$ and a classification head produces $\hat{z}_i = h_c(\mathbf{e}_i) \in \Delta^{K-1}$. The regression loss is $L_r(\theta) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (y_i - \hat{y}_i)^2$. The thresholds $\boldsymbol{\tau}$ partition $[-1, 1]$ into K intervals:

$$I_k(\boldsymbol{\tau}) = \begin{cases} [-1, \tau_1], & k = 1, \\ (\tau_{k-1}, \tau_k], & k = 2, \ldots, K-1, \\ (\tau_{K-1}, 1], & k = K. \end{cases} \tag{2}$$

Soft Assignment and Differentiability. We use a sigmoid-based soft assignment so the classification loss is differentiable in $\boldsymbol{\tau}$:

$$\pi_{ik}(\boldsymbol{\tau}) = \sigma\Big(\frac{y_i - \tau_{k-1}}{\gamma}\Big) - \sigma\Big(\frac{y_i - \tau_k}{\gamma}\Big), \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid, $\gamma > 0$ is a temperature, and $\tau_0 = -1$, $\tau_K = 1$ are fixed boundaries. Normalizing as $\pi_{ik} \leftarrow \pi_{ik} / \sum_{k'} \pi_{ik'}$; by telescoping, $\sum_k \pi_{ik}$ is $\boldsymbol{\tau}$-independent, and normalization therefore rescales each sample's gradient by a $\boldsymbol{\tau}$-independent (sample-dependent) factor and preserves the per-sample direction (Appendix A.4). As $\gamma \to 0$, the soft assignment recovers the hard indicator. The temperature γ is annealed via a cosine schedule from $\gamma_{\max}$ to $\gamma_{\min} > 0$.

Differentiable Category Statistics. The soft membership π_{ik} yields differentiable category statistics. The soft cardinality and soft mean are:

$$\widetilde{n}_k = \sum_i \pi_{ik}, \quad \bar{y}_k = \frac{\sum_i \pi_{ik} y_i}{\widetilde{n}_k + \epsilon}, \tag{4}$$

and the soft within-category variance is:

$$\widetilde{\mathrm{Var}}_k = \frac{\sum_i \pi_{ik} (y_i - \bar{y}_k)^2}{\widetilde{n}_k + \epsilon}, \tag{5}$$

all differentiable in $\boldsymbol{\tau}$ through π_{ik}. The auxiliary classification loss uses soft membership as targets:

$$L_c(\theta, \boldsymbol{\tau}) = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{\tau}) \cdot \log \hat{z}_{i,k}, \tag{6}$$

differentiable in both θ (through $\hat{z}$) and $\boldsymbol{\tau}$ (through π); the thresholds therefore receive gradient from both L_c and $\mathcal{R}$ and converge to partitions that are simultaneously discriminative and balanced.

Threshold Regularization. The regularizer $\mathcal{R}(\boldsymbol{\tau})$ prevents degenerate configurations through three complementary terms. Within-stratum variance encourages homogeneous sentiment within each category:

$$\mathcal{R}_{\text{var}}(\boldsymbol{\tau}) = \sum_{k=1}^{K} \frac{\widetilde{n}_k}{N} \widetilde{\text{Var}}_k. \tag{7}$$

Balance penalizes deviation of category sizes from $1/K$, preventing the auxiliary task from inheriting the marginal skew:

$$\mathcal{R}_{\text{bal}}(\boldsymbol{\tau}) = \lambda_1 \sum_{k=1}^{K} \Big(\frac{\widetilde{n}_k}{N} - \frac{1}{K}\Big)^2. \tag{8}$$

Spacing encourages uniform interval widths $\Delta_k = |I_k|$:

$$\mathcal{R}_{\text{space}}(\boldsymbol{\tau}) = \lambda_2 \sum_{k=1}^{K} \Big(\Delta_k - \frac{2}{K}\Big)^2. \tag{9}$$

The complete regularizer is $\mathcal{R}(\boldsymbol{\tau}) = \mathcal{R}_{\text{var}} + \mathcal{R}_{\text{bal}} + \mathcal{R}_{\text{space}}$, with λ_1, λ_2 trading off balance and spacing uniformity (their relative magnitude determines whether the learned partition is closer to quantile-based or uniform spacing).

Reparameterization and Initialization. Strict ordering $-1 < \tau_1 < \cdots < \tau_{K-1} < 1$ is enforced via unconstrained $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K) \in \mathbb{R}^K$:

$$\tau_k = -1 + 2 \cdot \frac{\sum_{j=1}^{k} \text{softplus}(\delta_j)}{\sum_{j=1}^{K} \text{softplus}(\delta_j)}, \quad k = 1, \ldots, K-1, \tag{10}$$

where $\text{softplus}(x) > 0$ guarantees monotonicity by construction. Thresholds are initialized at the empirical quantiles, $\tau_k^{(0)} = Q_{k/K}(\{y_i\})$.

4.3 Trajectory-Based Convergence Balancing

To address the task-level challenge, we propose TCB, as illustrated in Figure 4 (right). It tracks each task's convergence velocity along EMA loss trajectories, shifts the task weights $\lambda_r(t), \lambda_c(t)$ toward the slower task, and reweights the auxiliary categories in proportion to their losses.

Training Trajectory and Convergence Velocity. Smoothed loss trajectories come from exponential moving averages (EMAs) of each task:

$$\mu_r^t = \beta\, \mu_r^{t-1} + (1-\beta)\, L_r^t, \qquad \mu_c^t = \beta\, \mu_c^{t-1} + (1-\beta)\, L_c^t, \tag{11}$$

where $\beta \in (0,1)$ has effective lookback $W = \lfloor 1/(1-\beta) \rfloor$. Per-task convergence velocity is the clamped normalized rate of loss reduction over W steps:

$$\rho_r^t = \text{clamp}\Big(\frac{\mu_r^{t-W} - \mu_r^t}{\max(\mu_r^{t-W}, \epsilon)},\ -1,\ 1\Big), \qquad \rho_c^t = \text{clamp}\Big(\frac{\mu_c^{t-W} - \mu_c^t}{\max(\mu_c^{t-W}, \epsilon)},\ -1,\ 1\Big). \tag{12}$$

Inverse Progress Weighting. Unlike GradNorm (Chen et al., 2018) (instantaneous gradient norms), uncertainty weighting (Kendall et al., 2018) (static homoscedastic uncertainty), or dynamic weight averaging (Liu et al., 2019a) (single-step loss ratios), TCB maintains a windowed velocity estimate over W steps and applies inverse-progress weighting via softmax over negative velocities:

$$\lambda_r(t),\ \lambda_c(t) = \text{softmax}\Big(\frac{-\rho_r^t}{T_\lambda},\ \frac{-\rho_c^t}{T_\lambda}\Big), \tag{13}$$

where $T_\lambda > 0$ controls sensitivity. As the velocity gap widens, weights shift smoothly toward the lagging task. TCB requires no separate per-task backward passes (unlike GradNorm/PCGrad). During warm-up ($t < W$), $\lambda_r = \lambda_c = 0.5$.

Class-Level Reweighting. To address residual within-partition difficulty heterogeneity (§3), TCB additionally applies softmax-over-loss class reweighting; this reuses the loss-statistics infrastructure

of $\lambda_r(t), \lambda_c(t)$. Per-category loss $L^t_{c,k} = \frac{1}{|\mathcal{B}^t|} \sum_{i \in \mathcal{B}^t} \pi_{ik}(-\log \hat{z}_{i,k})$ satisfies $\sum_k L^t_{c,k} = L^t_c$. With $\mathcal{K}^t_{\mathcal{B}} = \{k : \sum_i \pi_{ik} > \epsilon\}$ the active categories,

$$q^t_k = \frac{\exp(L^t_{c,k}/T_q)}{\sum_{j \in \mathcal{K}^t_{\mathcal{B}}} \exp(L^t_{c,j}/T_q)}, \quad k \in \mathcal{K}^t_{\mathcal{B}}, \tag{14}$$

where $T_q > 0$ controls focus intensity (smaller T_q concentrates weight on the highest-loss category).

4.4 End-to-End Integration

End-to-End Training. Incorporating the trajectory-based task weights and the class-level reweighting into Eq. 1, the complete training objective of SAS (Figure 4, bottom) becomes:

$$\mathcal{L}(\theta, \boldsymbol{\tau}, t) = \lambda_r(t) \cdot L_r(\theta) + \lambda_c(t) \cdot \sum_{k \in \mathcal{K}^t_{\mathcal{B}}} q^t_k \, L_{c,k}(\theta, \boldsymbol{\tau}) + \mathcal{R}(\boldsymbol{\tau}), \tag{15}$$

where $\lambda_r(t), \lambda_c(t)$ are the task weights (Eq. 13), q^t_k the class weights (Eq. 14), and $\mathcal{R}(\boldsymbol{\tau})$ the threshold regularizer (Eqs. 7–9), kept outside $\lambda_c(t)$ to act throughout training. A single backward pass updates θ and $\boldsymbol{\delta}$ jointly; $\lambda_r(t), \lambda_c(t), q^t_k$ are stop-gradient, giving $\nabla_\theta \mathcal{L} = \lambda_r(t) \nabla_\theta L_r + \lambda_c(t) \sum_k q^t_k \nabla_\theta L_{c,k}$. Algorithm 1 (Appendix A.4) lists the complete procedure.

Framework Summary. SAS makes the label partition a trained parameter: DSP learns where the categories lie, TCB decides when their supervision matters, in one loss and one backward pass. The auxiliary head is discarded after training, so structured supervision adds no inference cost.

5 Experiments

5.1 Implementation details

Hardware and software. All runs use a single NVIDIA A100 (40 GB) with PyTorch 2.3.1 / CUDA 12.4 and AdamW with backbone learning rate 1×10^{-5}, SAS-module rate 1×10^{-3}, batch size 5 to 10 (memory-bound for 3B models), 100 epochs with cosine decay and 100 warmup steps. SAS defaults: K=5, T_λ=T_q=1.0, β=0.9. Details in Appendix A.9 and A.13.

Datasets. Ten regression datasets in three regimes: three private financial datasets (NEU, FXE, INV) on EUR/USD news and commentary, labelled in their source studies by a fine-tuned RoBERTa-Large model; three public financial benchmarks (FiQA-2018 (Maia et al., 2018); SemEval-2017 Task 5 Headline and Microblog (Cortis et al., 2017)); four cross-domain benchmarks (SST (Socher et al., 2013), CMU-MOSEI (Bagher Zadeh et al., 2018), CMU-MOSI (Zadeh et al., 2016), FB-VA (Preoţiuc-Pietro et al., 2016)). Targets are rescaled to $[-1, 1]$; private datasets use a 9:1 train/val split, and public splits follow Appendix A.8.

Backbones. Seven backbones, all fine-tuned end-to-end: BERT-base, BERT-large (Devlin et al., 2019), RoBERTa-base (Liu et al., 2019b), DeBERTa-v3-large (He et al., 2023), Llama-3.2-1B, Llama-3.2-3B (Llama Team, 2024), Qwen2.5-1.5B (Qwen Team, 2024). The main tables use RoBERTa-base, DeBERTa-v3-large and Llama-3.2-3B. Details in Appendix A.14.

Baselines. Ten methods in two families: *imbalanced regression* (LDS, FDS (Yang et al., 2021), Balanced-MSE (Ren et al., 2022), Ordinal-Entropy (Zhang et al., 2023), equal-width and quantile binning); *multi-task balancing* (UW (Kendall et al., 2018), GradNorm (Chen et al., 2018), DWA (Liu et al., 2019a), FAMO (Liu et al., 2023)). Single-task and Fixed-MTL (w_c=0.5) complete the twelve baselines. Reproductions match published numbers within 5%. Details in Appendix A.12.

Metrics. Five per cell: main-task MSE↓, MAE↓, Pearson r↑ and auxiliary K=5-class Acc↑ (%) and macro-F1↑ (%). Multi-seed means are reported for the main scaling, twelve-baseline, cross-domain, and component ablation experiments; the remaining secondary analyses report point estimates. Rows shared across tables (Single-task, Fixed-MTL, SAS on $\text{RoBERTa}_{\text{base}}$) report the same means. Standard deviations on multi-seed cells are within 0.0008 MSE; significance, where reported, uses Holm-Bonferroni paired t-tests at $p < 0.05$. Details in Appendix A.9 and A.12.

5.2 Main Results

Scaling across backbones. Table 1 compares SAS with Single-task and Fixed-MTL on NEU, FXE and INV across three backbone scales (RoBERTa-base, DeBERTa-v3-large, Llama-3.2-3B).

Table 1: Performance of SAS and Fixed-MTL on three private financial datasets (NEU/FXE/INV) across three backbone scales. MSE↓, MAE↓, Pearson r↑. Bold: SAS.

		NEU			FXE			INV		
Backbone	**Method**	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
$\text{RoBERTa}_{\text{base}}$	Single-task	0.0206	0.1200	0.886	0.0217	0.1122	0.814	0.0199	0.1144	0.884
	Fixed-MTL	0.0216	0.1223	0.881	0.0225	0.1239	0.809	0.0220	0.1184	0.865
	SAS (ours)	**0.0193**	**0.1073**	**0.892**	**0.0203**	**0.1120**	**0.828**	**0.0188**	**0.1110**	**0.886**
$\text{DeBERTa}_{\text{large}}$	Single-task	0.0194	0.1075	0.893	0.0204	0.1146	0.831	0.0188	0.1132	0.885
	Fixed-MTL	0.0203	0.1132	0.889	0.0212	0.1169	0.822	0.0197	0.1070	0.880
	SAS (ours)	**0.0183**	**0.1055**	**0.901**	**0.0191**	**0.1071**	**0.844**	**0.0177**	**0.1059**	**0.894**
$\text{Llama-3.2}_{\text{3B}}$	Single-task	0.0170	0.1003	0.905	0.0179	0.1108	0.850	0.0164	0.1050	0.899
	Fixed-MTL	0.0178	0.1061	0.901	0.0186	0.1036	0.842	0.0172	0.1006	0.898
	SAS (ours)	**0.0160**	**0.0990**	**0.906**	**0.0168**	**0.1015**	**0.856**	**0.0155**	**0.1006**	**0.910**

Table 2: Comparison of twelve baselines and SAS on private financial datasets ($\text{RoBERTa}_{\text{base}}$); MAE in Appendix A.12. Pink: best per column; blue: second best; gray rows: simple baselines. Italicized rows: closest competitors (DWA, FAMO). "n/a" = no auxiliary head.

	NEU				FXE				INV			
Method	MSE↓	r↑	Acc↑	F1↑	MSE↓	r↑	Acc↑	F1↑	MSE↓	r↑	Acc↑	F1↑
Single-task	0.0206	0.886	n/a	n/a	0.0217	0.814	n/a	n/a	0.0199	0.884	n/a	n/a
Fixed-MTL	0.0216	0.881	76.20	61.50	0.0225	0.809	75.20	59.90	0.0220	0.865	76.70	62.40
Quantile-bin	0.0207	0.883	76.30	62.20	0.0219	0.816	75.40	60.60	0.0210	0.869	77.20	63.50
Equal-width-bin	0.0205	0.884	76.20	61.90	0.0216	0.813	75.20	60.20	0.0208	0.872	77.10	63.20
LDS (Yang et al., 2021)	0.0201	0.893	77.30	62.90	0.0210	0.826	76.20	61.30	0.0209	0.876	78.20	64.30
FDS (Yang et al., 2021)	0.0204	0.889	77.00	63.00	0.0213	0.830	76.10	61.30	0.0203	0.879	78.10	64.20
Balanced-MSE (Ren et al., 2022)	0.0197	0.887	77.20	63.40	0.0212	0.824	76.10	61.80	0.0206	0.874	78.00	64.70
Ordinal-Entropy (Zhang et al., 2023)	0.0206	0.885	77.90	63.60	0.0217	0.819	77.10	61.70	0.0204	0.875	78.90	64.80
GradNorm (Chen et al., 2018)	0.0199	0.888	76.10	62.30	0.0210	0.822	75.30	60.50	0.0204	0.877	76.90	63.60
UW (Kendall et al., 2018)	0.0205	0.884	76.60	62.30	0.0213	0.821	75.80	60.50	0.0211	0.873	77.80	63.70
DWA (Liu et al., 2019a)	0.0195	0.891	77.10	62.60	0.0206	0.825	76.30	60.80	0.0198	0.881	78.20	64.10
FAMO (Liu et al., 2023)	0.0198	0.887	76.40	62.40	0.0209	0.820	75.60	60.70	0.0196	0.883	77.60	63.80
SAS (ours)	0.0193	0.892	79.60	66.40	0.0203	0.828	78.40	64.40	0.0188	0.886	80.20	67.20

SAS has the lowest MSE in all nine cells (0.0193 vs. 0.0206 for Single-task and 0.0216 for Fixed-MTL on NEU/RoBERTa-base), reduces MSE by 10.5% and MAE by 5.9% on average over Fixed-MTL, and raises Pearson r in every cell; on the auxiliary task it gains 3.5 pp Acc and 4.8 pp F1 (Table 28). Two patterns in the table locate the source of this gain. First, an auxiliary head that is added without structure or scheduling hurts: Fixed-MTL is worse than Single-task in all nine cells, the task-level interference identified in the motivation analysis, and SAS turns the same head into a gain by learning its categories and scheduling its weight. Second, a 24-fold increase in backbone size lowers the MSE of every method but leaves the relative reduction of SAS nearly unchanged (9.9% to 11.6%): additional capacity does not supply the missing category structure; the deficit SAS repairs thus lies in the supervision signal rather than in the representation. Details in Appendix A.14.

Comparison against twelve baselines. Table 2 compares SAS with twelve baselines under identical protocols on RoBERTa-base. SAS has the lowest MSE and MAE and the highest Acc and F1 on all three datasets; the closest competitor trails on MSE by 0.0002 to 0.0008 (e.g., DWA 0.0195 vs. SAS 0.0193 on NEU; FAMO 0.0196 vs. 0.0188 on INV). The baselines split along the two challenges of the motivation analysis. The imbalanced-regression methods (LDS, FDS, Balanced-MSE) reweight samples by label density: they lower MSE but raise Acc/F1 by only 1 to 2 pp over Fixed-MTL, because reweighting changes how much each sample counts without giving the auxiliary task categories that follow the skew, confirming that reweighting alone does not repair the structural deficit. DWA and FAMO come closest on MSE because they schedule the task weights, but they schedule a fixed partition; SAS exceeds them by 2 to 4 pp on Acc/F1, and the ablation attributes this margin to the learned partition, since trajectory scheduling on a frozen partition reaches 78.40 Acc on NEU and adding DSP raises it to 79.60 (Table 4). On Pearson r, SAS is best on INV and within 0.002 of LDS (NEU) and FDS (FXE); this small rank-correlation cost reflects the reallocation of error reduction toward the sparse extreme regions (Table 6). Details in Appendix A.12 and A.15.

Cross-domain generalization. On SST, CMU-MOSEI and CMU-MOSI, SAS lowers MSE by 9.4% to 10.6% over Fixed-MTL (Table 3) and wins all six backbone-by-dataset cells, with F1-macro $+3$ to $+4$ pp over the second-best baseline (LDS or DWA). These corpora share no domain with EUR/USD commentary but share its label geometry: continuous targets concentrated around a

neutral center, with imbalance ratios from about $2\times$ (CMU-MOSI) upward (Appendix A.8). SAS operates only on that geometry, as DSP partitions the label axis and TCB schedules the resulting task without domain-specific input; an equal gain on unrelated text thus shows that the method responds to label geometry rather than to financial content. Details in Appendix A.11 and A.17.

Public financial benchmarks. On the human-annotated FiQA-2018 and SemEval-2017 Task 5, SAS lowers MSE by 9.5% to 10.7% over Fixed-MTL on the canonical test splits (Table 3), with Acc $+3$ to $+4$ pp and F1 $+4$ to $+5$ pp. This setting separates the method from two properties of the private data. The private scores come from a fine-tuned RoBERTa-Large annotator, and SAS might be learning that annotator's own partition of the label space; with human labels the gain stays at the same size, which ties it to the score distribution rather than to how the scores were produced. The public test splits are also evaluated once, after model selection on a separate carve-out of the training data (Appendix A.9), and the gain holds on held-out data. Details in Appendix A.16.

Table 3: MSE on public financial and cross-domain datasets (RoBERTa$_{\text{base}}$); Δ vs. Fixed-MTL.

Block	Dataset	Fixed-MTL	SAS	Δ
Public financial	FiQA-2018	0.0612	**0.0548**	-10.5%
	SemEval Headline	0.1050	**0.0950**	-9.5%
	SemEval Microblog	0.0920	**0.0822**	-10.7%
Cross-domain	SST	0.0712	**0.0642**	-9.8%
	CMU-MOSEI	0.0808	**0.0722**	-10.6%
	CMU-MOSI	0.0874	**0.0792**	-9.4%

5.3 Ablation Study and Analysis

Component ablation. Table 4 removes each component on NEU/RoBERTa-base. DSP alone (fixed $w_c{=}0.5$) lowers MSE from 0.0216 to 0.0204, 52% of the joint gain, but leaves Acc/F1 at the Fixed-MTL level (76.20/61.50); TCB alone (frozen Fixed-MTL partition) lowers MSE to 0.0209 (30%) and lifts Acc/F1 to 78.40/64.80. Together they reach 0.0193 and 79.60/66.40, more than the sum of the single-component gains on MSE, MAE, Acc and F1. The two variants fail in complementary ways: with DSP alone, the early-converging classification task (Figure 3, middle) keeps a fixed gradient share and its accuracy stays flat; with TCB alone, the frozen categories cannot follow the regression error, capping the MSE gain at 30%. Only a schedule on a learned partition improves both metrics, the coupling argued in the method section.

Table 4: Component ablation on NEU (RoBERTa$_{\text{base}}$).

Variant	MSE↓	MAE↓	r↑	Acc↑	F1↑
Full SAS (DSP+TCB)	**0.0193**	**0.1073**	**0.892**	**79.60**	**66.40**
−DSP (TCB only)	0.0209	0.1206	0.885	78.40	64.80
−TCB (DSP only)	0.0204	0.1161	0.888	76.20	61.50
−DSP−TCB (Fixed-MTL)	0.0216	0.1223	0.881	76.20	61.50

Boundary regimes. SAS approaches Fixed-MTL where its design predicts (Table 5): a $K{=}2$ partition coarsens DSP's auxiliary targets, balanced subsamples remove the imbalance SAS targets, and small datasets ($N{<}500$) make DSP's soft cardinalities noisy. The margin narrows to 1.8% to 3.0% but stays positive: each regime removes one ingredient of the mechanism, and SAS falls back toward its Fixed-MTL backbone rather than below it. These three boundaries form a contract: SAS targets continuous regression with structured imbalance, $K \geq 3$ partitions, and $N \geq 500$.

Table 5: Boundary regimes on FXE and NEU (MSE).

Regime	Fixed-MTL	SAS	Δ
FXE			
$K{=}2$ Binary	0.0218	**0.0213**	-2.3%
Balanced	0.0225	**0.0221**	-1.8%
Small $N{<}500$	0.0243	**0.0236**	-2.9%
NEU			
$K{=}2$ Binary	0.0209	**0.0204**	-2.4%
Balanced	0.0216	**0.0212**	-1.9%
Small $N{<}500$	0.0233	**0.0226**	-3.0%

Per-region MSE compression. Table 6 decomposes the regression error on NEU/RoBERTa-base over five of the seven regions of the motivation analysis (all seven in Appendix A.18). Fixed-MTL shows the U-curve of neutral collapse, with extreme-to-neutral MSE ratio $3.85\times$. SAS cuts extreme-region MSE by about 30% (0.0443 to 0.0304 and 0.0462 to 0.0321) while neutral-region MSE rises from 0.0120 to 0.0165, compressing the ratio to $1.95\times$ at a lower overall MSE (0.0216 to 0.0193). The neutral region absorbs this cost because the balance term of DSP gives it one of K comparable categories instead of most of the supervision, and capacity spent on the dense center is redirected to the tails; this is the deliberate error reallocation the structural mechanism is designed to produce. The extreme-cell r rises by up to $+0.18$, and the lift grows monotonically with distance from neutral,

Table 6: Per-region decomposition on NEU/RoBERTa$_{\text{base}}$.

	Fixed-MTL		SAS	
Region	MSE↓	r ↑	MSE↓	r ↑
Negative Extreme	0.0443	0.678	**0.0304**	**0.851**
Negative Strong	0.0287	0.802	**0.0222**	**0.894**
Neutral	0.0120	0.920	0.0165	0.917
Positive Strong	0.0306	0.794	**0.0214**	**0.897**
Positive Extreme	0.0462	0.666	**0.0321**	**0.845**

Table 7: Imbalance-scaling sweep on NEU/RoBERTa$_{\text{base}}$.

Imbalance	**Method**	MSE↓	MAE↓	r ↑
r=1.5×	Fixed-MTL	0.0234	0.1221	0.867
	SAS	**0.0216**	**0.1173**	**0.878**
r=6×	Fixed-MTL	0.0350	0.1493	0.793
	SAS	**0.0245**	**0.1249**	**0.861**
r=20×	Fixed-MTL	0.0450	0.1693	0.724
	SAS	**0.0270**	**0.1311**	**0.845**

exactly as the structured-supervision hypothesis predicts: the improvement is largest where the label density is lowest, which a uniform regularization effect would not produce (Appendix A.18).

Advantage scales with imbalance severity. Table 7 sweeps the imbalance ratio r, the neutral-to-extreme count ratio, on resampled NEU. From r=1.5 to r=20, Fixed-MTL MSE rises by 92% (0.0234 to 0.0450) and Single-task MSE by 80% (Appendix A.19), while SAS rises by 25% (0.0216 to 0.0270), a 3.2× to 3.7× slower degradation; the margin widens rather than saturates at extreme imbalance. The baselines degrade because the extremes' loss share shrinks with their counts. DSP breaks this link: the balance term keeps the extreme categories at a comparable share of the auxiliary supervision whatever r is, keeping per-region gradients visible, while TCB reallocates gradient toward the slower task as imbalance widens the convergence gap. A widening margin is the signature of a structural remedy: reweighting scales its correction with the observed counts, whereas the learned partition fixes the tails' share independently of them. Details in Appendix A.19.

6 Related Work

Imbalanced Regression. Density-based methods reweight the loss: LDS/FDS (Yang et al., 2021) smooth label density and feature statistics, Balanced MSE (Ren et al., 2022) corrects the label prior, and DenseLoss (Steininger et al., 2021) weights by inverse density, extending focal, class-balanced and cost-sensitive losses (Lin et al., 2017; Cui et al., 2019; Araf et al., 2024). Discretization methods classify over balanced target quantizations at several granularities (Xiong & Yao, 2024), pair soft-labeled group classification with within-group regression (Pu et al., 2025), learn thresholds over fixed ranks (Cao et al., 2020), or regularize features with ordinal entropy (Zhang et al., 2023). Both lines fix the label-space structure before training; SAS learns the partition jointly with the model.

Multi-Task Weighting. Adaptive methods weight tasks by training signals: gradient magnitudes in GradNorm (Chen et al., 2018), learned uncertainty in UW (Kendall et al., 2018), consecutive loss ratios in DWA (Liu et al., 2019a), and per-step loss decrease in FAMO (Liu et al., 2023), while PCGrad (Yu et al., 2020) projects away conflicting gradients. Such instantaneous signals miss a persistent convergence asymmetry; CoBa (Gong et al., 2024) tracks validation-loss slopes. TCB uses windowed training-loss velocities and schedules the task whose categories DSP learns.

Financial Sentiment Analysis. The field has moved from deep networks (Correia et al., 2022; Aslam et al., 2022) to LLM analyzers (Krugmann & Hartmann, 2024; Carneros-Prado et al., 2023). Its corpora are neutral-dominated (Omarkhan et al., 2021; Davidovic & McCleary, 2025), and existing remedies, augmentation (Assunção et al., 2024) and threshold adjustment (Leevy et al., 2023), act on discrete labels. SAS models it as continuous regression with learned label structure.

7 Conclusion

We identify two coupled challenges in financial sentiment regression, skew-induced neutral collapse and auxiliary task interference and convergence asymmetry, and propose SAS: DSP learns the auxiliary label partition, and TCB schedules its supervision from windowed convergence velocities in one loss. On three financial datasets and three backbones, SAS lowers MSE by 10.5% over Fixed-MTL, and the gain extends to seven public and cross-domain datasets. On NEU/RoBERTa-base it compresses the extreme-to-neutral MSE ratio from 3.85× to 1.95×, and under 20× skew its MSE rises by 25% versus 92% for Fixed-MTL. This widening margin marks a structural remedy: the missing category structure is learned with the model and scheduled by its convergence trajectory.

REFERENCES

Ahmed Alsayat. Improving sentiment analysis for social media applications using an ensemble deep learning language model. *Arabian Journal for Science and Engineering*, 47(2):2499–2511, 2022.

Imane Araf, Ali Idri, and Ikram Chairi. Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57(4):80, 2024.

Muhammad Asif, Atiab Ishtiaq, Haseeb Ahmad, Hanan Aljuaid, and Jalal Shah. Sentiment analysis of extremism in social media from textual information. *Telematics and Informatics*, 48:101345, 2020.

Naila Aslam, Furqan Rustam, Ernesto Lee, Patrick Bernard Washington, and Imran Ashraf. Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. *IEEE Access*, 10:39313–39324, 2022. doi: 10.1109/ACCESS.2022.3165621.

Gabriel O Assunção, Rafael Izbicki, and Marcos O Prates. Is augmentation effective in improving prediction in imbalanced datasets? *Journal of Data Science*, 2024.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

N. L. Balasudarsun, Bikramaditya Ghosh, and Sathish Mahendran. Impact of negative tweets on diverse assets during stressful events: An investigation through time-varying connectedness. *Journal of Risk and Financial Management*, 15(6):260, 2022.

Kousik Barik and Sanjay Misra. Analysis of customer reviews with an improved vader lexicon classifier. *Journal of Big Data*, 11(1):10, 2024.

Staphord Bengesi, Timothy Oladunni, Ruth Olusegun, and Halima Audu. A machine learning-sentiment analysis on monkeypox outbreak: An extensive dataset to show the polarity of public opinion from twitter tweets. *IEEE Access*, 11:11811–11826, 2023.

Paula Branco, Luís Torgo, and Rita P. Ribeiro. A survey of predictive modeling on imbalanced domains. *ACM Computing Surveys*, 49(2):1–50, 2016. doi: 10.1145/2907070.

Gregorius Satia Budhi, Raymond Chiong, Ilung Pranata, and Zhongyi Hu. Using machine learning to predict the sentiment of online reviews: a new framework for comparative analysis. *Archives of Computational Methods in Engineering*, 28:2543–2566, 2021.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020. doi: 10.1016/j.patrec.2020.11.008.

David Carneros-Prado, Laura Villa, Esperanza Johnson, Cosmin C Dobrescu, Alfonso Barragán, and Beatriz García-Martínez. Comparative study of large language models as emotion and sentiment analysis systems: A case-specific analysis of gpt vs. ibm watson. In *International Conference on Ubiquitous Computing and Ambient Intelligence*, pp. 229–239. Springer, 2023.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pp. 794–803. PMLR, 2018.

Filipe Correia, Ana Maria Madureira, and Jorge Bernardino. Deep neural networks applied to stock market sentiment analysis. *Sensors*, 22(12):4409, 2022.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Hürlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, 2017. doi: 10.18653/v1/s17-2089.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Milivoje Davidovic and Jacqueline McCleary. News sentiment and stock market dynamics: A machine learning investigation. *Journal of Risk and Financial Management*, 18(8):412, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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 (NAACL-HLT)*, pp. 4171–4186, 2019.

Hongcheng Ding, Xiangyu Shi, Ruiting Deng, Salaar Faroog, Deshinta Arrova Dewi, Shamsul Nahar Abdullah, and Bahiah A Malek. Eur/usd exchange rate forecasting incorporating text mining based on pre-trained language models and deep learning methods. *arXiv preprint arXiv:2411.07560*, 2024a.

Hongcheng Ding, Xuanze Zhao, Ruiting Deng, Shamsul Nahar Abdullah, and Deshinta Arrova Dewi. Eur-usd exchange rate forecasting based on information fusion with large language models and deep learning methods. *arXiv preprint arXiv:2408.13214*, 2024b.

Hongcheng Ding, Xuanze Zhao, Ruiting Deng, Shamsul Nahar Abdullah, Deshinta Arrova Dewi, and Zixiao Jiang. Dynamic adaptive optimization for effective sentiment analysis fine-tuning on large language models. *arXiv preprint arXiv:2408.11856*, 2024c.

Benjamin Garner, Corliss Thornton, Anita Luo Pawluk, Roberto Mora Cortez, Wesley Johnston, and Cesar Ayala. Utilizing text-mining to explore consumer happiness within tourism destinations. *Journal of Business Research*, 139:1366–1377, 2022.

Zi Gong, Hang Yu, Cong Liao, Bingchang Liu, Chaoyu Chen, and Jianguo Li. Coba: Convergence balancer for multitask finetuning of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8063–8077, 2024.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Rashid Khan, Furqan Rustam, Khadija Kanwal, Arif Mehmood, and Gyu Sang Choi. Us based covid-19 tweets sentiment analysis using textblob and supervised machine learning algorithms. In *2021 international conference on artificial intelligence (ICAI)*, pp. 1–8. IEEE, 2021.

Jan Ole Krugmann and Jochen Hartmann. Sentiment analysis in the age of generative ai. *Customer Needs and Solutions*, 11(1):3, 2024.

Joffrey L Leevy, Justin M Johnson, John Hancock, and Taghi M Khoshgoftaar. Threshold optimization and random undersampling for imbalanced credit card data. *Journal of Big Data*, 10(1):58, 2023.

Xuerong Li, Wei Shang, and Shouyang Wang. Text-based crude oil price forecasting: A deep learning approach. *International Journal of Forecasting*, 35(4):1548–1560, 2019.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Bo Liu, Yihao Feng, Peter Stone, and Qiang Liu. Famo: Fast adaptive multitask optimization. *Advances in Neural Information Processing Systems*, 36, 2023.

Jingfang Liu, Yingyi Zhou, Xiaoyan Jiang, and Wei Zhang. Consumers' satisfaction factors mining and sentiment analysis of b2c online pharmacy reviews. *BMC Medical Informatics and Decision Making*, 20(1):194, 2020.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1871–1880, 2019a.

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

Llama Team. The Llama 3 herd of models, 2024. arXiv:2407.21783. https://ai.meta.com/research/publications/the-llama-3-herd-of-models/.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1608.03983.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018 (WWW '18)*, pp. 1941–1942, 2018.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796, 2014. doi: 10.1002/asi.23062.

Robert Marcec and Robert Likic. Using twitter for sentiment analysis towards astrazeneca/oxford, pfizer/biontech and moderna covid-19 vaccines. *Postgraduate medical journal*, 98(1161):544–550, 2022.

Gaurav Meena, Krishna Kumar Mohbey, and Ajay Indian. Categorizing sentiment polarities in social networks data using convolutional neural network. *SN Computer Science*, 3(2):116, 2022.

Gonçalo Azevedo Mendes and Bruno Martins. Quantifying valence and arousal in text with multilingual pre-trained transformers. In *Advances in Information Retrieval (ECIR 2023)*, volume 13980, pp. 84–100. Springer, 2023.

M. Mujahid, E. Kına, F. Rustam, et al. Data oversampling and imbalanced datasets: an investigation of performance for machine learning and feature engineering. *Journal of Big Data*, 11:87, 2024. doi: 10.1186/s40537-024-00943-4.

A Naresh and P Venkata Krishna. An efficient approach for sentiment analysis using machine learning algorithm. *Evolutionary intelligence*, 14(2):725–731, 2021.

Usman Naseem, Imran Razzak, Katarzyna Musial, and Muhammad Imran. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. *Future Generation Computer Systems*, 113:58–69, 2020.

Zineb Nassr, Faouzia Benabbou, Nawal Sael, and Touria Hamim. Improving sentiment analysis performance on imbalanced moroccan dialect datasets using resample and feature extraction techniques. *Information*, 16(1):39, 2025.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-task learning as a bargaining game. In *International Conference on Machine Learning*, pp. 16428–16446, 2022.

Mahmood Mohammadi Nezhad, Saeed Rouhani, Navid Mohammadi, and Ali Shahedi. A unified GRU model for cryptocurrency price prediction and harsh price movement detection using enhanced sentiment analysis. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-46271-w.

Moldir Omarkhan, Gulnur Kissymova, and Iskander Akhmetov. Handling data imbalance using cnn and lstm in financial news sentiment analysis. In *2021 16th international conference on electronics computer and computation (ICECCO)*, pp. 1–8. IEEE, 2021.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. Modelling valence and arousal in facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pp. 9–15, 2016.

Ruizhi Pu, Gezheng Xu, Ruiyi Fang, Bing-Kun Bao, Charles Ling, and Boyu Wang. Leveraging group classification with descending soft labeling for deep imbalanced regression. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(19):19978–19985, 2025. doi: 10.1609/aaai.v39i19.34200.

Cheng Qian, Nitya Mathur, Nor Hidayati Zakaria, Rameshwar Arora, Vedika Gupta, and Mazlan Ali. Understanding public opinions on social media for financial sentiment analysis using ai-based techniques. *Information Processing & Management*, 59(6):103098, 2022.

Qwen Team. Qwen2.5 technical report, 2024. Technical report; `https://arxiv.org/abs/2412.15115`.

Golshid Ranibaran, Mohammad-Shahram Moin, Sasan H Alizadeh, and Abbas Koochari. Analyzing effect of news polarity on stock market prediction: a machine learning approach. In *2021 12th International Conference on Information and Knowledge Technology (IKT)*, pp. 102–106. IEEE, 2021.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for imbalanced visual regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7926–7935, 2022.

Loukas Samaras, Elena García-Barriocanal, and Miguel-Angel Sicilia. Sentiment analysis of covid-19 cases in greece using twitter data. *Expert Systems with Applications*, 230:120577, 2023.

Raj Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. When FLUE meets FLANG: Benchmarks and large pretrained language model for financial domain. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2322–2335, 2022.

Zhiqi Shao, Xusheng Yao, Feng Chen, Ze Wang, and Junbin Gao. Revisiting time-varying dynamics in stock market forecasting: A multi-source sentiment analysis approach with large language model. *Decision Support Systems*, 190:114362, 2025. doi: 10.1016/j.dss.2024.114362.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Michael Steininger, Konstantin Kobs, Padraig Davidson, Anna Krause, and Andreas Hotho. Density-based weighting for imbalanced regression. *Machine Learning*, 110:2187–2211, 2021.

Paul C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3):1139–1168, 2007. doi: 10.1111/j.1540-6261.2007.01232.x.

Paul C. Tetlock, Maytal Saar-Tsechansky, and Sofus Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The Journal of Finance*, 63(3):1437–1467, 2008. doi: 10.1111/j.1540-6261.2008.01362.x.

Sandra Wankmüller and Christian Heumann. How to estimate continuous sentiments from texts using binary training data. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pp. 182–192, 2021.

Zhanming Wen, Yanjun Chen, Hongwei Liu, and Zhouyang Liang. Text mining based approach for customer sentiment and product competitiveness using composite online review data. *Journal of Theoretical and Applied Electronic Commerce Research*, 19(3):1776–1792, 2024.

Haipeng Xiong and Angela Yao. Deep imbalanced regression via hierarchical classification adjustment. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. doi: 10.1109/cvpr52733.2024.02239.

Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 11842–11851. PMLR, 2021.

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

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016.

Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. In *International Conference on Learning Representations*, 2023.

Huiliang Zhao, Zhenghong Liu, Xuemei Yao, and Qin Yang. A machine learning-based sentiment analysis of online product reviews with a novel term weighting and feature selection approach. *Information Processing & Management*, 58(5):102656, 2021.

Muhammad Zulqarnain, Rozaida Ghazali, Muhammad Aamir, and Yana Mazwin Mohmad Hassim. An efficient two-state gru based on feature attention mechanism for sentiment analysis. *Multimedia Tools and Applications*, 83(1):3085–3110, 2024.

AI USE STATEMENT

We used generative AI tools to translate the text (Claude Opus 5.0), to check grammar and spelling and polish the wording (GPT-5.6), to guide the drawing of the figures (Gemini), and to check the content (Grok). We did not use generative AI tools to modify code. We have reviewed all AI-assisted work: we checked every reference for hallucinated citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The details needed to reproduce the results of this paper, including the datasets and their score normalization, the backbones and hyperparameters, the training procedure, the evaluation metrics, and the software and hardware setup, are disclosed in this appendix.

APPENDIX

Table 8 lists the contents of each part of the appendix.

APPENDIX A.0: LIMITATIONS

Hyperparameter sensitivity. SAS adds a partition count K, two softmax temperatures for task and class weighting, a momentum for the loss trajectories, and weights for the threshold regularizer. These values decide how finely the label space is divided and how sharply supervision is redistributed, and a practitioner moving SAS to a new dataset needs to know how its error responds when they depart from the defaults. The paper lacks a systematic sensitivity analysis over these hyperparameters; its main comparisons use one default configuration.

Table 8: What each part of the appendix holds.

Section	Contents
Appendix A.0	Limitations
Appendix A.1	Extended related work
Appendix A.2	Distributional skew and neutral collapse analysis
Appendix A.3	Task-level challenge analysis
Appendix A.4	SAS joint optimization and training procedure
Appendix A.5	DSP implementation details
Appendix A.6	Evaluation metrics
Appendix A.7	Benchmark configurations
Appendix A.8	Datasets and score normalization
Appendix A.9	Hyperparameters and environment
Appendix A.10	Backbone selection rationale
Appendix A.11	Cross-domain validation experiments
Appendix A.12	Baseline comparison
Appendix A.13	Computational cost
Appendix A.14	Backbone scaling on private financial datasets
Appendix A.15	Twelve-baseline comparison, detailed analysis
Appendix A.16	Public financial benchmark generalization
Appendix A.17	Cross-domain generalization, detailed analysis
Appendix A.18	Per-region MSE decomposition, detailed analysis
Appendix A.19	Imbalance-severity sweep, detailed analysis
Appendix A.20	Component ablation, detailed analysis
Appendix A.21	Boundary conditions, detailed analysis

Robustness to label noise and distribution shift. The private scores come from a fine-tuned annotator model, and financial text drifts over time; training labels can therefore be noisy, and the test distribution can depart from the training one. DSP learns category boundaries from the observed labels, and errors in those labels or shifts in their distribution act directly on the partition that SAS supervises. The paper lacks a controlled evaluation of SAS under injected label noise and under train–test label shift.

Design choices within each component. The ablation removes DSP or TCB as a whole, which shows that both components are needed but not which of their internal choices carry the gain. Within DSP, the spacing regularizer encourages uniform interval widths; within TCB, the momentum of the loss trajectories smooths the convergence-velocity estimates. Isolating such choices tells a reader which parts of SAS can be simplified without loss, and the paper lacks this finer-grained ablation.

APPENDIX A.1: EXTENDED RELATED WORK

Sentiment Analysis. Sentiment analysis is widely used across domains. In finance, it improves forecasting accuracy for crude oil prices (Li et al., 2019) and stock market movements (Correia et al., 2022), and the analysis of NFT-related tweets correlates public sentiment with market trends (Qian et al., 2022). Beyond finance, text mining of online reviews assesses product competitiveness (Wen et al., 2024) and the factors that influence consumer happiness in travel experiences (Garner et al., 2022).

Lexicon-based Methods. Lexicon-based methods rely on predefined sentiment dictionaries and face limitations from that reliance, yet they are applied across many fields: multi-domain customer reviews (Barik & Misra, 2024), online pharmacy reviews (Liu et al., 2020), and public sentiment on social media during the COVID-19 pandemic (Khan et al., 2021; Marcec & Likic, 2022; Samaras et al., 2023).

Traditional Methods. Classical machine-learning algorithms such as SVM, Random Forest and Naïve Bayes analyze public sentiment on disease outbreaks (Bengesi et al., 2023), support stock price prediction (Ranibaran et al., 2021), and classify multilingual extremism text (Asif et al., 2020), while a hybrid algorithm targets Twitter sentiment (Naresh & Venkata Krishna, 2021) and a machine-learning framework predicts the sentiment of online reviews (Budhi et al., 2021). Neural models

extend this line with LSIBA-ENN (Zhao et al., 2021), CNNs and LSTMs (Meena et al., 2022), a two-state GRU with feature attention (Zulqarnain et al., 2024), an ensemble of FastText and LSTM (Alsayat, 2022), the transformer-based DICET for Twitter (Naseem et al., 2020), and an LSTM-GRU ensemble for cryptocurrency-related tweets (Aslam et al., 2022).

Imbalanced Regression. Imbalanced learning has long been studied for classification, where resampling and cost adjustment rebalance minority classes (He & Garcia, 2009; Mujahid et al., 2024); predictive modeling on imbalanced domains extends the problem to continuous targets, where rarity is defined by a relevance function over the label range (Branco et al., 2016). For deep regression, LDS (Yang et al., 2021) smooths the empirical label density with a symmetric kernel and FDS smooths feature statistics across neighboring targets, Balanced MSE (Ren et al., 2022) corrects the mismatch between the training label prior and a balanced one, and DenseLoss (Steininger et al., 2021) weights each sample by the inverse of its estimated density. These methods carry the logic of focal loss (Lin et al., 2017), class-balanced loss (Cui et al., 2019), and cost-sensitive learning (Araf et al., 2024) from classes to a continuum: they change how much each sample counts, and the label space stays unstructured.

Discretization and Ordinal Structure. A complementary line gives the regression target discrete structure. Hierarchical classification adjustment (Xiong & Yao, 2024) classifies over balanced target quantizations at several granularities, and group classification (Pu et al., 2025) pairs group classification under descending soft labels with within-group regression. Ordinal regression learns rank-consistent thresholds over a fixed set of ranks (Cao et al., 2020), and ordinal entropy (Zhang et al., 2023) regularizes regression features toward the higher entropy of classification features while preserving label order. In each case the discrete structure is fixed before training, by quantization, grouping, or given ranks. SAS learns the partition boundaries jointly with the model; equal-width and quantile binning correspond to its frozen-threshold variants.

Multi-Task Weighting. Multi-task learning shares one representation across objectives and must decide how much each objective contributes. Loss-based schemes set the weights from optimization signals: GradNorm (Chen et al., 2018) tunes weights so that gradient norms track each task's relative training rate, uncertainty weighting (Kendall et al., 2018) scales each loss by a learned homoscedastic noise term, DWA (Liu et al., 2019a) weights tasks by the ratio of their losses in consecutive epochs, and FAMO (Liu et al., 2023) balances the per-step decrease of the task losses. Gradient-based schemes resolve conflicts directly: PCGrad (Yu et al., 2020) projects each task gradient off conflicting directions, and Nash-MTL (Navon et al., 2022) casts the joint update as a bargaining game; both require per-task gradients. For LLM fine-tuning, CoBa (Gong et al., 2024) balances convergence using the slopes of validation losses over a history window. TCB shares this convergence view, computes windowed velocities from training losses, weights tasks by a softmax of negative velocities, and adds class-level reweighting within the auxiliary task, whose categories DSP learns.

Financial Sentiment under Imbalance. Text-based measures of media tone and firm fundamentals (Tetlock, 2007; Tetlock et al., 2008) and annotated corpora such as the Financial PhraseBank (Malo et al., 2014) established financial sentiment as a learning problem, and LLMs are now used directly as sentiment and emotion analyzers (Krugmann & Hartmann, 2024; Carneros-Prado et al., 2023). Financial sentiment data are dominated by neutral content (Omarkhan et al., 2021; Davidovic & McCleary, 2025), and sentiment studies handle skewed labels with recurrent and convolutional architectures (Omarkhan et al., 2021), data augmentation (Assunção et al., 2024), resampling (Nassr et al., 2025), or threshold optimization with undersampling (Leevy et al., 2023). These remedies act on discrete labels. SAS models financial sentiment as continuous regression and addresses the skew through learned label-space structure.

APPENDIX A.2: DISTRIBUTIONAL SKEW AND NEUTRAL COLLAPSE ANALYSIS

This appendix provides detailed analysis of the data distribution imbalance and neutral collapse phenomenon discussed in Section 2. Each panel of Figure 2 is presented individually with extended discussion.

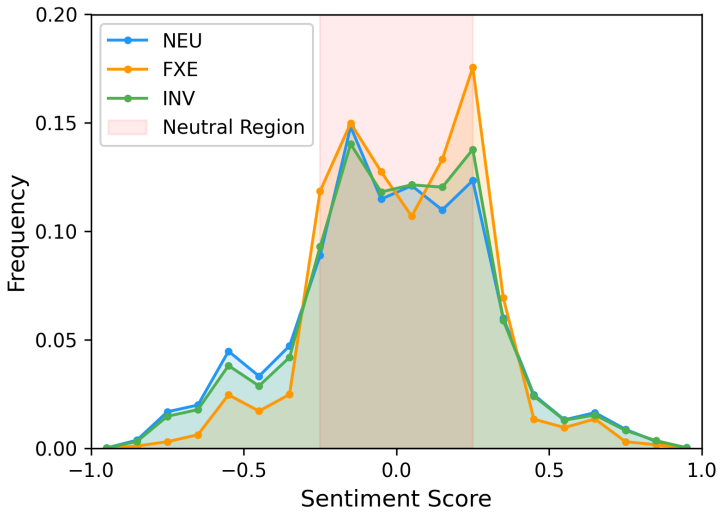


Figure 5: Training data distributions across three financial sentiment datasets (NEU, FXE, INV). All three datasets exhibit extreme concentration in the neutral region ($|y| < 0.25$), with over 60% of samples falling within this range. Extreme sentiments ($|y| > 0.6$) account for fewer than 7% of samples. The shaded neutral region highlights the dense core that dominates the MSE loss landscape during fine-tuning.

The training distributions of all three financial sentiment datasets share a characteristic pattern: a pronounced unimodal concentration around the neutral region. The NEU dataset shows peak frequencies of approximately 15% in the $[-0.20, -0.10)$ range, FXE peaks at 17.6% in the $[0.20, 0.30)$ range, and INV shows a broader plateau from $[-0.20, 0.30)$ with frequencies around 12–14%. Despite differences in skewness direction across datasets, the common feature is severe underrepresentation of extreme sentiments: scores beyond ± 0.6 collectively account for fewer than 7% of samples in each dataset. This distributional pattern reflects the inherent characteristics of financial news, where most articles convey moderate sentiment while strongly positive or negative events are relatively rare.

When a standard LLM is fine-tuned on these skewed distributions using MSE loss, the resulting predictions collapse toward the neutral region. The model predictions show extreme concentration around the $[-0.2, 0.2]$ range, with peak frequencies exceeding 16%. The model demonstrates virtually no capability to predict extreme sentiments, with predictions rarely extending beyond ± 0.4. This prediction pattern indicates that the model has learned to minimize expected loss by defaulting to safe, neutral values, as the high density of near-zero training samples overwhelms the gradient signal from sparse extreme samples.

To quantify the neutral collapse effect, per-region MSE is computed across seven fine-grained sentiment regions: $[-1, -0.6)$, $[-0.6, -0.3)$, $[-0.3, -0.1)$, $[-0.1, 0.1)$, $[0.1, 0.3)$, $[0.3, 0.6)$, and $[0.6, 1]$. The results reveal a clear U-shaped pattern: MSE increases monotonically with distance from the neutral center. Under Fixed-MTL, the extreme regions exhibit MSE approximately $3.85\times$ higher than the neutral region; SAS compresses this ratio to $1.95\times$ (Table 6). The near-neutral regions $[-0.3, -0.1)$ and $[0.1, 0.3)$ show intermediate error levels of 0.019–0.020, while the moderate regions $[-0.6, -0.3)$ and $[0.3, 0.6)$ exhibit errors of 0.029–0.031. This fine-grained analysis demonstrates that neutral collapse is not a binary extreme-vs-neutral phenomenon, but a continuous gradient of degradation that scales with distance from the neutral center.

Appendix A.3: Task-Level Challenge Analysis

This appendix provides detailed analysis of the task-level optimization challenges discussed in Section 2. Each panel of Figure 3 is presented individually with extended discussion.

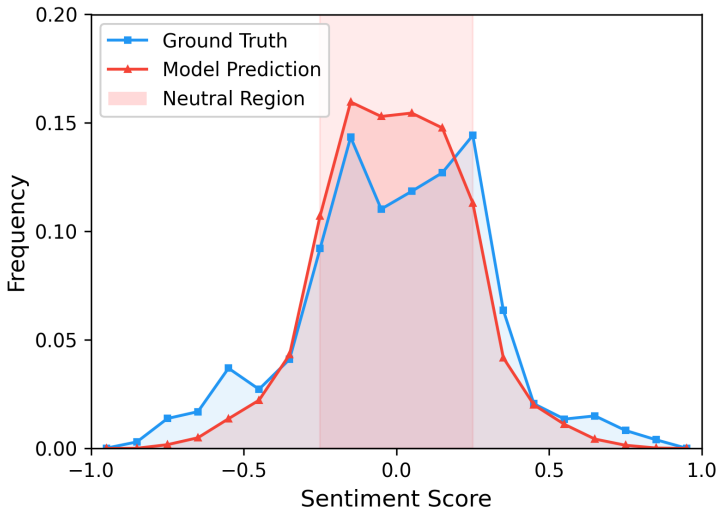


Figure 6: Comparison between ground-truth test set distribution and model predictions from a fine-tuned LLM. The prediction distribution is dramatically narrower than the ground truth, with density concentrating near zero while the tails are severely under-covered. This is the neutral collapse phenomenon: the model learns to predict safe, central values rather than capturing the full range of sentiment intensity.

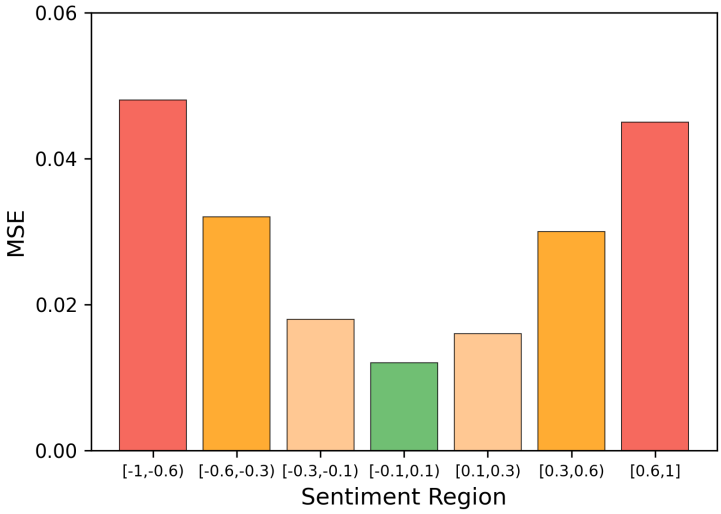


Figure 7: Per-region MSE analysis across seven sentiment regions from extreme negative to extreme positive. Prediction error in extreme regions is up to 3.85× higher than in the neutral region under Fixed-MTL, quantifying the severity of neutral collapse.

A model is trained with loss $L = (1-w_c)L_r + w_c L_c$ for fixed values of $w_c \in \{0.00, 0.05, \ldots, 0.50\}$. At $w_c = 0$ (pure regression), the model achieves an MSE of 0.0273 (higher than the tuned Single-task baseline in Table 2, since the sweep holds one shared recipe fixed across all w_c rather than re-tuning each point) and no classification accuracy exists since no classification head is trained. As w_c increases, the auxiliary classification signal provides structural supervision that improves regression performance, with MSE reaching its minimum of approximately 0.0194 around $w_c = 0.25$. Beyond this point, the classification objective begins to dominate, degrading regression performance. Meanwhile, classification accuracy follows a different trajectory, peaking at approximately 77.7% around

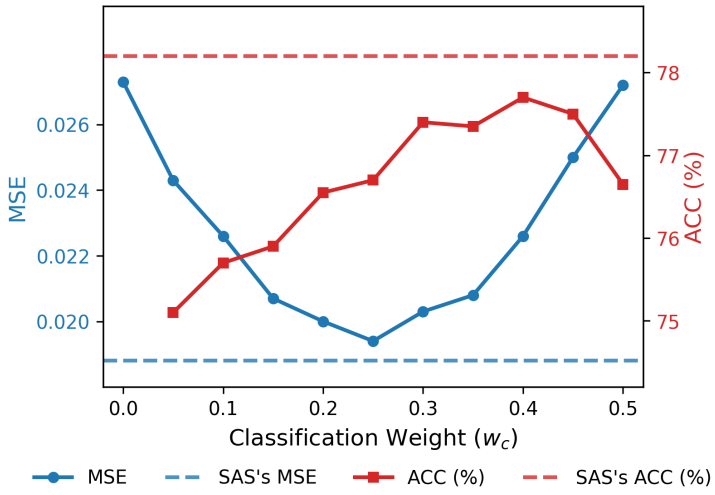


Figure 8: Sensitivity of MSE and ACC to fixed classification weight w_c. MSE (blue) exhibits a U-shape: decreasing as w_c increases from 0 to 0.25, then increasing as w_c exceeds 0.30. ACC (red) shows an inverted-U, peaking around $w_c = 0.40$. The MSE-optimal and ACC-optimal weights differ, demonstrating that no single fixed weight simultaneously optimizes both objectives. Dashed lines indicate the performance of SAS, which dominates all fixed-weight configurations on both metrics.

$w_c = 0.40$. This misalignment between the MSE-optimal weight ($w_c \approx 0.25$) and the ACC-optimal weight ($w_c \approx 0.40$) demonstrates the inherent tension in fixed-weight multi-task learning. SAS, shown as dashed lines, simultaneously outperforms all fixed-weight configurations on both metrics.

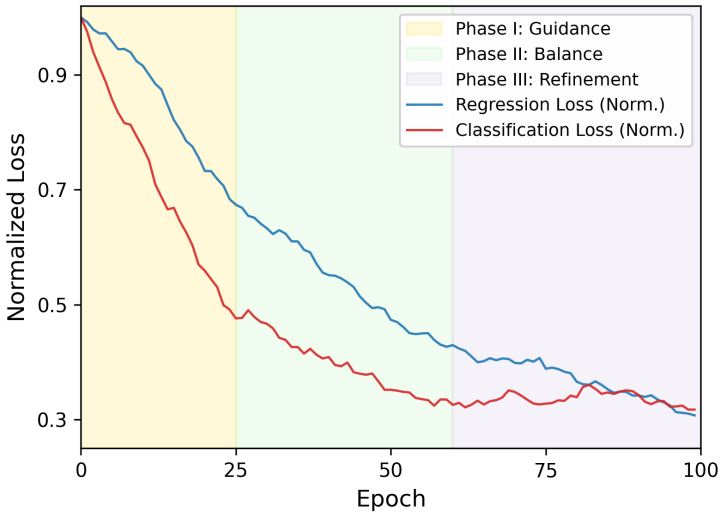


Figure 9: Normalized loss trajectories for regression (blue) and classification (red) over 100 training epochs. The two tasks exhibit pronounced convergence-velocity asymmetry: classification converges rapidly early (epochs 0–25), velocities equalize in the middle range (epochs 25–60), and classification plateaus while regression continues to descend in the late range (epochs 60–100). A fixed weight cannot track this changing velocity ratio.

Fixed weights fail for a reason visible in the normalized loss trajectories of both tasks over training. Each loss is normalized by its initial value at epoch 0, enabling direct comparison of convergence

rates despite different absolute scales. The two trajectories exhibit pronounced velocity asymmetry. In the early range (epochs 0–25), the classification loss drops rapidly from 1.0 to approximately 0.45, while the regression loss decreases more gradually to around 0.65. This asymmetry reflects the coarser nature of classification: discrete category boundaries are easier to learn than continuous value regression. In the middle range (epochs 25–60), the per-task convergence velocities equalize and both tasks descend at comparable rates. In the late range (epochs 60–100), classification plateaus around 0.35–0.40 while regression continues to decrease, eventually reaching approximately 0.30, and the velocity ratio inverts. A single fixed weight cannot match this time-varying velocity profile: a setting tuned to the early classification-dominant range becomes suboptimal in the late regression-dominant range. This time-varying velocity ratio motivates an adaptive scheme that responds to the observed convergence trajectory itself, which is the design of TCB (§4.3).

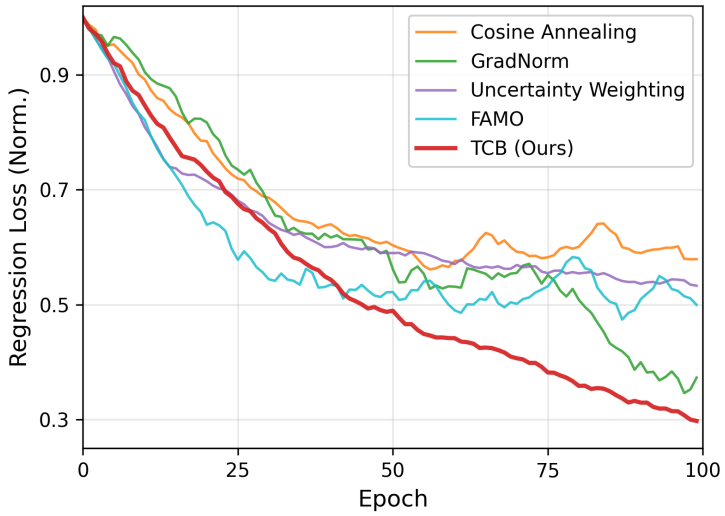


Figure 10: Comparison of dynamic weighting methods on regression loss convergence. Cosine annealing (orange) exhibits late-stage rebounds. GradNorm (green) oscillates throughout training. Uncertainty weighting (purple) plateaus prematurely. FAMO (cyan) converges fast initially but becomes unstable in mid-to-late training. TCB (red) achieves the lowest and most stable convergence.

Beyond fixed weights, existing dynamic weighting methods also leave this challenge unresolved. Five approaches are compared on regression loss convergence: (1) Cosine annealing (Loshchilov & Hutter, 2017) follows a predetermined schedule that decays the classification weight over training. While it performs reasonably in early epochs, it cannot adapt to the non-monotonic dynamics observed in the loss trajectories, causing visible loss rebounds around epochs 65 and 82 when the schedule mismatches the actual training dynamics. (2) GradNorm (Chen et al., 2018) adjusts weights based on gradient magnitudes to balance task learning rates. However, in the two-task setting here, the gradient signals are noisy, leading to oscillatory weight updates that propagate into unstable loss trajectories throughout training. (3) Uncertainty weighting (Kendall et al., 2018) learns task-specific uncertainty parameters to weight each loss. While it converges quickly in early epochs, it plateaus prematurely around a normalized loss of 0.35, as it misinterprets low task uncertainty as convergence rather than recognizing the need for continued optimization. (4) FAMO (Liu et al., 2023) employs a fast adaptive optimization strategy that achieves rapid initial convergence but exhibits significant instability in mid-to-late training, with multiple oscillation episodes around epochs 35, 55, and 78, as its approximate optimization becomes unreliable when the loss landscape becomes non-convex. (5) TCB tracks per-task convergence velocity over a windowed exponential moving average and applies softmax-of-negative-velocity to redirect gradient toward the slower-converging task, achieving the lowest final loss of approximately 0.25 with smooth, stable convergence across the early, middle, and late training ranges.

APPENDIX A.4: SAS JOINT OPTIMIZATION AND TRAINING PROCEDURE

This appendix provides the implementation details that complement the main-text §4.4 description: a derivation supporting the soft-assignment normalization claim, the mini-batch implementation of the differentiable category statistics, the stability mechanisms used during joint optimization of $\boldsymbol{\tau}$ and θ, the computational overhead introduced by SAS over standard single-task training, and the complete end-to-end training procedure summarized in Algorithm 1.

Soft-Assignment Normalization. We verify that the normalization $\pi_{ik} \leftarrow \pi_{ik}/Z_i$ with $Z_i = \sum_{k=1}^{K} \pi_{ik}$ rescales the gradient by a $\boldsymbol{\tau}$-independent factor. Recall $\pi_{ik}(\boldsymbol{\tau}) = \sigma((y_i - \tau_{k-1})/\gamma) - \sigma((y_i - \tau_k)/\gamma)$. Summing over k telescopes:

$$\begin{aligned} \sum_{k=1}^{K} \pi_{ik}(\boldsymbol{\tau}) &= \sigma\left(\frac{y_i - \tau_0}{\gamma}\right) - \sigma\left(\frac{y_i - \tau_K}{\gamma}\right) \\ &= \sigma\left(\frac{y_i + 1}{\gamma}\right) - \sigma\left(\frac{y_i - 1}{\gamma}\right), \end{aligned} \tag{16}$$

where the second equality uses the fixed boundaries $\tau_0 = -1$ and $\tau_K = 1$. Hence $\partial Z_i/\partial \tau_k = 0$ for all $k = 1, \ldots, K-1$, and $\partial(\pi_{ik}/Z_i)/\partial \tau_k = (\partial \pi_{ik}/\partial \tau_k)/Z_i$. The factor $1/Z_i$ depends on the sample (through y_i) but not on $\boldsymbol{\tau}$; thus normalization preserves the per-sample gradient direction with respect to $\boldsymbol{\tau}$.

Differentiable Category Statistics: Implementation. The sums $\sum_{i=1}^{N}$ in Eqs. 4–5 are evaluated on the current mini-batch $\mathcal{B}^t$ (replacing N with $|\mathcal{B}^t|$) for stochastic optimization. When $|\mathcal{B}^t|$ is small relative to K, an EMA running estimate over batches stabilizes $\widetilde{n}_k$, $\bar{y}_k$, and $\widetilde{\mathrm{Var}}_k$.

Joint Optimization Stability. Since the thresholds $\boldsymbol{\tau}$ and model parameters θ are updated jointly, changes to $\boldsymbol{\tau}$ alter the classification targets π_{ik} for all samples. To prevent oscillation, a dedicated learning rate is used for the threshold parameters $\boldsymbol{\delta}$ (1×10^{-3}, matched to the SAS-module rate), and the γ-annealing schedule ensures that early in training, when thresholds are still adjusting, the soft assignments change gradually. Empirically, the learned thresholds converge within the first few epochs and remain stable thereafter.

Computational Overhead. The additional cost per step consists of computing per-class soft cross-entropy across $K{=}5$ partitions (vector ops), maintaining two scalar EMAs for trajectory tracking, and applying a softmax over K values for class reweighting. Unlike gradient-based methods such as GradNorm (Chen et al., 2018) and PCGrad (Yu et al., 2020) that require separate backward passes per task, SAS operates entirely on loss statistics and requires only a single backward pass. Empirically, total wall-clock training time on the same backbone (RoBERTa-base) increases by approximately 58% over Single-task and 41% over Fixed-MTL (Appendix A.13). The overhead is dominated by the DSP soft-cardinality computation; the TCB component itself adds negligible cost (scalar EMAs and a 2-element softmax).

Training Procedure. The complete training procedure is summarized in Algorithm 1. The SAS module attaches to any backbone by adding a classification head and the auxiliary loss term. The key hyperparameters are K (number of categories), λ_1, λ_2 (regularizer weights), β (EMA smoothing), and T_q (class reweighting temperature), while W is derived automatically from β and γ follows its preset cosine annealing schedule.

APPENDIX A.5: DSP IMPLEMENTATION DETAILS

This appendix details the DSP component introduced in Section 4.2. The DSP design uses three coupled mechanisms: softplus reparameterization to enforce strict threshold ordering, sigmoid-based soft assignment for differentiability, and quantile initialization. All threshold updates flow through standard backpropagation jointly with the model parameters, sharing a single AdamW optimizer rather than a separate solver.

Algorithm 1 SAS Framework: End-to-End Training

Require: Dataset $\mathcal{D}$, backbone $\mathcal{M}$, categories K, class temperature T_q, task temperature T_λ, EMA coefficient β, γ schedule $[\gamma_{\max}, \gamma_{\min}]$

Initialize thresholds from quantiles: $\tau_k^{(0)} = Q_{k/K}(\{y_i\})$ for $k = 1, \ldots, K-1$
Convert to unconstrained parameters: $\boldsymbol{\delta} \leftarrow \mathrm{softplus}^{-1}(\text{increments from } \boldsymbol{\tau}^{(0)})$
Set lookback window: $W \leftarrow \lfloor 1/(1-\beta) \rfloor$
Perform initial forward pass to obtain L_r^0 and L_c^0
Initialize EMAs: $\mu_r^0 \leftarrow L_r^0$, $\mu_c^0 \leftarrow L_c^0$
for each mini-batch $\mathcal{B}^t$ **do**
  Anneal γ^t via cosine schedule from $\gamma_{\max}$ to $\gamma_{\min}$
  Compute thresholds $\boldsymbol{\tau}$ from $\boldsymbol{\delta}$ via Eq. 10
  Compute soft assignments $\pi_{ik}(\boldsymbol{\tau})$ via Eq. 3, normalize $\pi_{ik} \leftarrow \pi_{ik} / \sum_{k'} \pi_{ik'}$
  Forward pass: compute $\hat{y}_i = h_r(\mathbf{e}_i)$, $\hat{z}_i = h_c(\mathbf{e}_i)$
  Compute L_r^t and per-class $L_{c,k}^t$ via Eq. 6
  Update EMAs μ_r^t, μ_c^t via Eq. 11
  if $t \geq W$ **then**
    Compute convergence rates ρ_r^t, ρ_c^t via Eq. 12
    Compute task weights $\lambda_r(t), \lambda_c(t)$ via Eq. 13
  else
    $\lambda_r(t), \lambda_c(t) \leftarrow 0.5, 0.5$ {equal weighting during warm-up}
  end if
  Compute class weights q_k^t via Eq. 14
  Compute regularizer $\mathcal{R}(\boldsymbol{\tau}) = \mathcal{R}_{\text{var}} + \mathcal{R}_{\text{bal}} + \mathcal{R}_{\text{space}}$ (Eq. 7–9)
  Compute total loss $\mathcal{L}(\theta, \boldsymbol{\tau}, t)$ via Eq. 15
  Single backward pass: update θ and $\boldsymbol{\delta}$ jointly
end for

A.5.1 Reparameterization, Soft Assignment, and Joint Optimization

Softplus reparameterization for strict ordering. The thresholds $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{K-1})$ must satisfy $-1 < \tau_1 < \cdots < \tau_{K-1} < 1$. Following Eq. 10, $\boldsymbol{\tau}$ is parameterized through unconstrained variables $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K) \in \mathbb{R}^K$ and the softplus function $\mathrm{softplus}(x) = \log(1 + e^x) > 0$:

$$\tau_k = -1 + 2 \cdot \frac{\sum_{j=1}^{k} \mathrm{softplus}(\delta_j)}{\sum_{j=1}^{K} \mathrm{softplus}(\delta_j)}, \quad k = 1, \ldots, K-1. \tag{17}$$

Because every increment $\mathrm{softplus}(\delta_j)$ is strictly positive, monotonicity holds by construction without explicit projection or barrier terms; no Lagrangian or projection step is ever invoked.

Soft cardinality and category statistics. Using the soft assignment $\pi_{ik}(\boldsymbol{\tau}) = \sigma((y_i - \tau_{k-1})/\gamma) - \sigma((y_i - \tau_k)/\gamma)$ from Eq. 3, the soft cardinality is

$$\widetilde{n}_k = \sum_{i \in \mathcal{B}^t} \pi_{ik}(\boldsymbol{\tau}), \tag{18}$$

evaluated on the current mini-batch. Because π_{ik} is differentiable in $\boldsymbol{\tau}$, both $\widetilde{n}_k$ and the soft mean / variance of Eq. 4–5 are differentiable, and the regularizer terms $\mathcal{R}_{\text{var}}, \mathcal{R}_{\text{bal}}, \mathcal{R}_{\text{space}}$ contribute gradient to $\boldsymbol{\delta}$ through standard backpropagation.

Joint backprop with AdamW. The complete loss $\mathcal{L}(\theta, \boldsymbol{\tau}, t)$ in Eq. 15 is differentiable jointly in the model parameters θ and the unconstrained threshold parameters $\boldsymbol{\delta}$. A single backward pass populates $\nabla_\theta \mathcal{L}$ and $\nabla_{\boldsymbol{\delta}} \mathcal{L}$. Both are passed to AdamW, with the SAS-module learning rate (1×10^{-3}) for $\boldsymbol{\delta}$, letting the partition adjust in step with the auxiliary head. Thresholds require no dedicated solver; the partition is updated in lockstep with the backbone over the same 100-epoch cosine schedule used everywhere else.

Regularizer weights. The regularizer $\mathcal{R}(\boldsymbol{\tau}) = \mathcal{R}_{\text{var}} + \mathcal{R}_{\text{bal}} + \mathcal{R}_{\text{space}}$ uses default $\lambda_1 = \lambda_2 = 0.01$ for the balance and spacing terms (Eq. 8–9); $\mathcal{R}_{\text{var}}$ has unit weight.

Number of categories K. The default is K=5, matching the granularity of standard fine-grained sentiment annotation (strong negative, negative, neutral, positive, strong positive).

A.5.2 QUANTILE INITIALIZATION

Initialization formula. The thresholds are initialized at the empirical quantiles of the training-set scores:

$$\tau_k^{(0)} = Q_{k/K}(\{y_i\}_{i=1}^N), \quad k = 1, \ldots, K-1, \tag{19}$$

where $Q_p(\cdot)$ denotes the p-th sample quantile. The unconstrained variables $\boldsymbol{\delta}$ are then set so that the softplus reparameterization recovers $\boldsymbol{\tau}^{(0)}$ exactly.

Why quantile. Quantile initialization places the partition at the (initially) optimal balanced configuration with respect to the marginal label distribution: each category contains approximately N/K samples at $t = 0$. The balance regularizer $\mathcal{R}_{\text{bal}}$ then exerts vanishing pressure during the first epochs because the partition is already near-balanced. As training proceeds, the within-stratum variance term $\mathcal{R}_{\text{var}}$ pulls the boundaries away from the marginal quantile and toward partitions that are coherent with respect to the embedding geometry, while $\mathcal{R}_{\text{space}}$ prevents collapse of any single interval. Empirically the thresholds converge within the first few epochs and remain stable thereafter (Appendix A.4).

Adaptive across distributions. Because the initialization is data-driven, it adapts automatically to datasets with different distributional skew (heavy-neutral private financial datasets, positively-skewed SemEval, broad-spectrum SST and MOSEI). This is the source of SAS's backbone-agnostic behavior across the ten datasets in §5: no per-dataset tuning of the partition is needed.

Interpretability of the learned thresholds. The converged partitions are directly readable as sentiment bands. On NEU (RoBERTa-base, K=5), the quantile initialization $\boldsymbol{\tau}^{(0)} = (-0.28, -0.06, 0.07, 0.30)$ converges to $\boldsymbol{\tau}^* = (-0.41, -0.09, 0.11, 0.44)$: the inner pair stays close to the quantile positions, preserving the boundary of the dense neutral band, while the within-stratum variance term pushes the outer pair outward toward wider mild-polarity intervals, reserving the two extreme categories for scores beyond roughly ± 0.4, where directional conviction is unambiguous. This learned geometry mirrors the region structure used in the per-region analysis (Appendix A.18) and explains where the extreme-region MSE compression originates.

APPENDIX A.6: EVALUATION METRICS

Regression Task. For the sentiment score regression task, where the model predicts a continuous sentiment polarity score, we employ the following metrics:

- **Mean Squared Error (MSE)** calculates the average squared differences between the predicted sentiment polarity scores ($\hat{y}_i$) and the annotated sentiment polarity scores (y_i) in the test set. MSE provides a measure of the model's accuracy in predicting the exact sentiment scores. It is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{20}$$

  where n is the number of texts in the test set.

- **Mean Absolute Error (MAE)** measures the average magnitude of the absolute errors between the predicted and annotated sentiment polarity scores, ignoring their direction. MAE helps in understanding the average error magnitude and is less sensitive to outliers compared to MSE. It is formulated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|\,. \tag{21}$$

- **Pearson Correlation Coefficient** (r) measures the linear correlation between predicted and ground-truth sentiment scores. Values close to 1 indicate strong positive correlation; values close to 0 indicate independence. It is defined as:

$$r = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{22}$$

  where $\bar{\hat{y}}$ and $\bar{y}$ denote the means of the predictions and ground-truth scores respectively.

Classification Task. For the sentiment classification task, where texts are categorized into five sentiment classes (e.g., strong positive, positive, neutral, negative, strong negative), we calculate the following metrics:

- **Accuracy (ACC)**: the ratio of correctly predicted sentiment classes to the total number of predictions. It measures the overall correctness of the model's classifications and is defined as:
$$\text{ACC} = \frac{\sum_{i=1}^{5} TP_i}{\sum_{i=1}^{5}(TP_i + FN_i)}, \tag{23}$$
where TP_i is the number of instances correctly predicted as class i, and FN_i is the number of instances that actually belong to class i but are wrongly predicted as other classes.
- **Precision**: the ratio of correctly predicted positive instances to all instances predicted as positive. It measures the model's ability to avoid false positives and is calculated as:
$$\text{Precision} = \frac{TP_{pos}}{TP_{pos} + FP_{pos}}, \tag{24}$$
where TP_{pos} is the number of positive instances correctly predicted as positive, and FP_{pos} is the number of non-positive instances wrongly predicted as positive.
- **Recall (or Sensitivity)**: the ratio of correctly predicted positive instances to all actual positive instances. It measures the model's ability to identify all positive instances and is calculated as:
$$\text{Recall} = \frac{TP_{pos}}{TP_{pos} + FN_{pos}}, \tag{25}$$
where FN_{pos} is the number of positive instances wrongly predicted as non-positive.
- **F1 Score**: the harmonic mean of Precision and Recall, providing a balanced measure of the model's performance, especially when the sentiment classes are imbalanced. It is calculated as:
$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{26}$$

Weighted recall equals accuracy. In a multi-class classification problem with n classes, let TP_i, FN_i, and Support_i denote the true positives, false negatives, and total instances (support) for class i, respectively. The recall for class i is given by:

$$\text{Recall}_i = \frac{TP_i}{TP_i + FN_i} = \frac{TP_i}{\text{Support}_i}. \tag{27}$$

The weighted recall is then calculated as:

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{n} \text{Recall}_i \times \text{Support}_i}{\sum_{i=1}^{n} \text{Support}_i}. \tag{28}$$

Substituting the expression for Recall_i:

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{n} \frac{TP_i}{\text{Support}_i} \times \text{Support}_i}{\sum_{i=1}^{n} \text{Support}_i}. \tag{29}$$

The Support_i terms in the numerator and denominator cancel out, leaving:

$$\text{Weighted Recall} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} \text{Support}_i}. \tag{30}$$

This is exactly the formula for overall accuracy:

$$\text{Accuracy} = \frac{TP_1 + TP_2 + \ldots + TP_n}{\text{Support}_1 + \text{Support}_2 + \ldots + \text{Support}_n}. \tag{31}$$

Therefore, in a multi-class classification setting, weighted recall is mathematically equivalent to accuracy.

APPENDIX A.7: BENCHMARK

We provide the full benchmark configurations in Table 9. B1 and B2 are single-task and Fixed-MTL baselines that quantify the neutral-collapse phenomenon. B3-SAS through B8-SAS instantiate SAS on six additional full fine-tuned backbones spanning encoder (BERT/DeBERTa) and decoder (Llama/Qwen) families, to verify generalizability across architectures and scales. All configurations use full parameter fine-tuning.

Table 9: Benchmark models and experimental settings (full fine-tuning throughout).

Benchmark	**Model**	**Task**
B1	RoBERTa-base (Liu et al., 2019b)	Single-task regression
B2	RoBERTa-base (Liu et al., 2019b)	Multi-task (Fixed-MTL, w_c=0.5)
B3-SAS	BERT-base (Devlin et al., 2019)	Multi-task with SAS (ours)
B4-SAS	BERT-large (Devlin et al., 2019)	Multi-task with SAS (ours)
B5-SAS	DeBERTa-v3-large (He et al., 2023)	Multi-task with SAS (ours)
B6-SAS	Llama-3.2-1B (Llama Team, 2024)	Multi-task with SAS (ours)
B7-SAS	Llama-3.2-3B (Llama Team, 2024)	Multi-task with SAS (ours)
B8-SAS	Qwen2.5-1.5B (Qwen Team, 2024)	Multi-task with SAS (ours)

APPENDIX A.8: DATASETS

A.8.1 OVERVIEW AND SCORE NORMALIZATION

We evaluate SAS on ten English-language sentiment regression datasets organized into three blocks: (i) three private financial datasets (NEU, FXE, INV), each comprising news and analytical commentary on EUR/USD exchange rates; (ii) three public financial benchmarks (FiQA-2018, SemEval-2017 Task 5 Microblog, SemEval-2017 Task 5 Headline); and (iii) four cross-domain benchmarks (continuous-variant SST, CMU-MOSEI text-only, CMU-MOSI text-only, FB-VA). All ten datasets carry continuous sentiment targets with non-trivial class imbalance, and together cover an imbalance spectrum from approximately $2.0\times$ (CMU-MOSI) to $29.6\times$ (FXE), enabling controlled study of neutral collapse across distributional regimes.

The financial block reflects the canonical setting where neutral sentiment dominates due to the cautious nature of market commentary (Li et al., 2019; Correia et al., 2022; Mujahid et al., 2024). Financial texts contain domain-specific terminology, implicit sentiments tied to market conditions, and complex semantic relationships (Omarkhan et al., 2021; Davidovic & McCleary, 2025), motivating dataset-aware regression rather than off-the-shelf classification. The cross-domain block stress-tests SAS under domain shifts spanning movie reviews, spoken-text opinion videos, and social-media valence-arousal annotations, ensuring that observed gains are not specific to financial discourse.

Score Normalization. Native target ranges differ across datasets: NEU/FXE/INV/FiQA/SemEval-2017 are already in $[-1, 1]$, SST in $[0, 1]$, CMU-MOSEI/CMU-MOSI in $[-3, 3]$, and FB-VA in $[1, 9]$. For cross-dataset comparability and consistent application of the SAS regression head (which targets $[-1, 1]$), we apply per-dataset affine normalization $y' = ay + b$ as listed in Table 10. Each transformation maps the dataset's natural neutral mid-point to 0 and the extremes to ± 1. The neutral region is uniformly defined as the normalized interval $[-0.25, 0.25]$, and all per-region statistics in this Appendix are reported on the normalized scale.

A.8.2 INV DATASET

The INV dataset (Ding et al., 2024a) captures investor sentiment in EUR/USD news and analyst commentary and contains 23,241 financial texts. Annotations are produced using a RoBERTa-Large model fine-tuned on financial sentiment, yielding scores in $[-1.000, 0.900]$ with mean -0.044 and standard deviation 0.303. The slight negative mean reflects the cautious tone characteristic of currency-market commentary, particularly during periods of macroeconomic uncertainty.

The INV distribution concentrates 67.4% of samples in the neutral region $[-0.25, 0.25]$, with imbalance ratio $9.33\times$ between the densest and sparsest regions. The mid-positive cluster (regions

Table 10: Score normalization mapping each native sentiment range to $[-1, 1]$ via affine transformation $y' = ay + b$. The neutral mid-point of every dataset is mapped to 0.

Dataset(s)	Native range	a	b	Transformation
NEU, FXE, INV, FiQA, SemEval-2017	$[-1, 1]$	1	0	$y' = y$ (identity)
SST (Socher et al., 2013)	$[0, 1]$	2	-1	$y' = 2y - 1$
CMU-MOSEI (Bagher Zadeh et al., 2018), CMU-MOSI (Zadeh et al., 2016)	$[-3, 3]$	$1/3$	0	$y' = y/3$
FB-VA (Preoţiuc-Pietro et al., 2016)	$[1, 9]$	$1/4$	$-5/4$	$y' = (y - 5)/4$

Table 11: Distribution of sentiment polarity scores in the INV dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	839	0.0361
2	$[-0.6, -0.3)$	2,534	0.1090
3	$[-0.3, -0.1)$	5,421	0.2333
4	$[-0.1, 0.1]$	5,571	0.2397
5	$(0.1, 0.3]$	6,002	0.2582
6	$(0.3, 0.6]$	2,231	0.0960
7	$(0.6, 1.0]$	643	0.0277

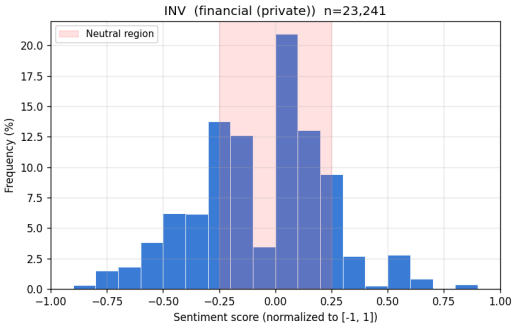


Figure 11: Distribution of sentiment polarity scores in the INV dataset.

4–5) reflects the dataset's investing.com source where buy-side analysis is over-represented, while both extremes are sparsely populated. This pattern motivates the structured partitioning of SAS to prevent the regression head from collapsing toward the dense centre.

A.8.3 NEU Dataset

The NEU dataset (Ding et al., 2024b) comprises 12,780 financial text samples extracted from investing.com and forexempire.com, focusing on EUR/USD exchange rate news articles and analytical content. Annotations are produced by a RoBERTa-Large model fine-tuned on financial sentiment, with scores in $[-0.950, 0.900]$, mean -0.060, and standard deviation 0.312. The slight negative mean is consistent with the negative-sentiment bias inherent in financial news discourse and market commentary.

The NEU distribution concentrates 65.5% of samples in the neutral region with imbalance ratio $8.21\times$. Regions 3–5 carry roughly equal mass (23–24% each), forming a wide neutral plateau that the SAS classification head must subdivide rather than treat as a single class. This is the canonical setting DSP targets.

A.8.4 FXE Dataset

The FXE dataset (Ding et al., 2024c) originates from an EUR/USD exchange rate forecasting study that integrates text mining with pre-trained language models, and comprises 10,461 currency-focused financial texts. Sentiment is annotated by RoBERTa-Large fine-tuned on the same currency domain, yielding scores in $[-0.900, 0.900]$ with mean -0.014 and standard deviation 0.258. The near-zero mean reflects a more balanced representation of bullish and bearish market views than NEU or INV.

FXE is the most imbalanced dataset in our lineup with imbalance ratio $29.63\times$ and 73.2% of samples falling in the neutral region. The combined mass of regions 3–5 exceeds 81%, while extreme regions

Table 12: Distribution of sentiment polarity scores in the NEU dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	523	0.0409
2	$[-0.6, -0.3)$	1,604	0.1255
3	$[-0.3, -0.1)$	3,031	0.2372
4	$[-0.1, 0.1]$	3,017	0.2361
5	$(0.1, 0.3]$	2,983	0.2334
6	$(0.3, 0.6]$	1,253	0.0980
7	$(0.6, 1.0]$	369	0.0289

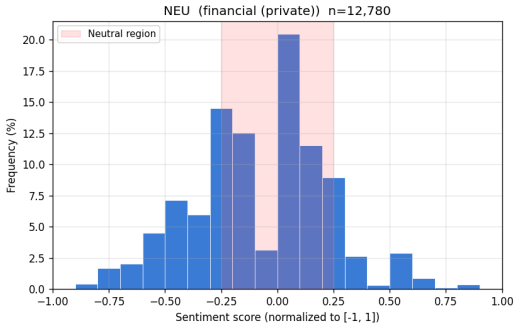


Figure 12: Distribution of sentiment polarity scores in the NEU dataset.

Table 13: Distribution of sentiment polarity scores in the FXE dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	109	0.0104
2	$[-0.6, -0.3)$	699	0.0668
3	$[-0.3, -0.1)$	2,809	0.2685
4	$[-0.1, 0.1]$	2,453	0.2345
5	$(0.1, 0.3]$	3,230	0.3088
6	$(0.3, 0.6]$	967	0.0924
7	$(0.6, 1.0]$	194	0.0185

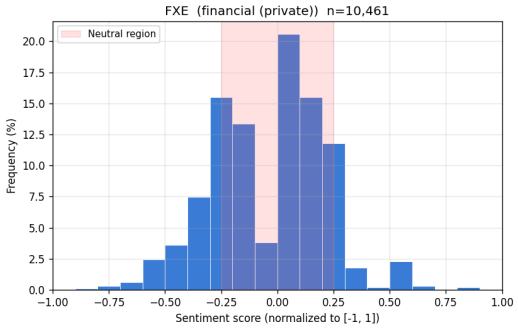


Figure 13: Distribution of sentiment polarity scores in the FXE dataset.

1 and 7 together account for less than 3%. This makes FXE the showcase setting for neutral collapse and the most stringent test of SAS's ability to recover extreme-region accuracy.

A.8.5 FiQA-2018 Dataset

FiQA-2018 (Maia et al., 2018) is the financial sentiment subtask of the WWW'18 Open Challenge on Financial Opinion Mining and Question Answering. It contains 1,173 short texts (microblog posts and news headlines) with continuous polarity scores in $[-0.940, 0.940]$ assigned by domain annotators per company-aspect. We follow the standard FLUE benchmark split of 822 train / 117 dev / 234 test (Shah et al., 2022). After identity normalization, scores have mean 0.122 and standard deviation 0.399.

FiQA exhibits a pronounced positive skew (region 6 alone accounts for 36.5%) with imbalance ratio $10.70\times$, the inverse of the NEU/FXE/INV pattern. Only 26.3% of samples fall in the neutral region. Including FiQA verifies that SAS does not require neutral-heavy distributions to function: the structured partitioning must instead concentrate capacity on the dense mid-positive cluster while preserving the sparse extremes. Reported MSE benchmarks on FiQA range from FLANG-ELECTRA at 0.034 to BERT-base at 0.073 (Shah et al., 2022), providing direct comparators for the SAS results.

A.8.6 SemEval-2017 Task 5 Microblog Dataset

The SemEval-2017 Task 5 Microblog subtask (5.1) (Cortis et al., 2017) contains 1,700 financial microblogs from Twitter and StockTwits, each annotated for bullish/bearish sentiment as a continuous score in $[-1, 1]$ with respect to a tagged company. The original task uses weighted cosine similarity

Table 14: Distribution of sentiment polarity scores in the FiQA-2018 dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	40	0.0341
2	$[-0.6, -0.3)$	224	0.1910
3	$[-0.3, -0.1)$	99	0.0844
4	$[-0.1, 0.1]$	91	0.0776
5	$(0.1, 0.3]$	211	0.1799
6	$(0.3, 0.6]$	428	0.3649
7	$(0.6, 1.0]$	80	0.0682

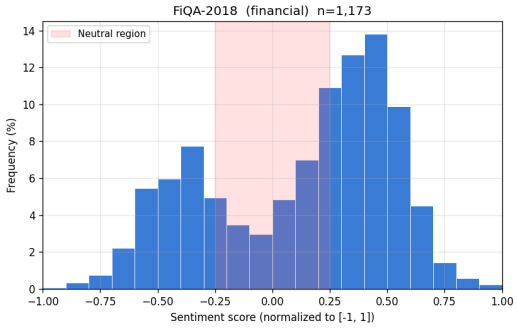


Figure 14: Distribution of sentiment polarity scores in the FiQA-2018 dataset.

as its primary metric; we report MSE, MAE, and Pearson r for cross-task comparability. Scores have mean 0.115 and standard deviation 0.380.

Table 15: Distribution of sentiment polarity scores in the SemEval-2017 Microblog dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	31	0.0182
2	$[-0.6, -0.3)$	376	0.2212
3	$[-0.3, -0.1)$	131	0.0771
4	$[-0.1, 0.1]$	122	0.0718
5	$(0.1, 0.3]$	268	0.1576
6	$(0.3, 0.6]$	697	0.4100
7	$(0.6, 1.0]$	75	0.0441

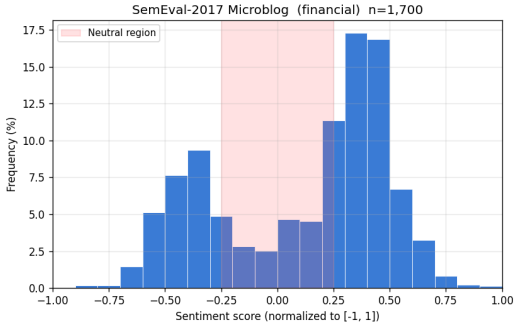


Figure 15: Distribution of sentiment polarity scores in the SemEval-2017 Microblog dataset.

SemEval-Microblog is the most imbalanced public dataset in our lineup with ratio $22.48\times$, second only to FXE. The mid-positive cluster (region 6) dominates at 41.0%, while the neutral region holds only 21.4%. This pattern is the inverse of the private financial datasets and tests whether SAS adapts to bullish-skewed distributions where extreme-positive recovery is less critical than mid-cluster discrimination.

A.8.7 SemEval-2017 Task 5 Headline Dataset

The SemEval-2017 Task 5 Headline subtask (5.2) (Cortis et al., 2017) comprises 1,142 financial news headlines (plus 491 test items) annotated with continuous sentiment scores in $[-1, 1]$ per tagged company entity. Scores have mean 0.031 and standard deviation 0.394, indicating a near-balanced sentiment distribution at the headline level.

Headline is the most balanced of the three public financial datasets with imbalance ratio $4.89\times$ and only 37.7% of samples in the neutral region. The relatively flat distribution across regions 2–6 makes Headline a moderate-imbalance test where SAS gains over fixed-weight multi-task baselines should be smaller, providing an upper-tail control on the improvement claims.

Table 16: Distribution of sentiment polarity scores in the SemEval-2017 Headline dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	81	0.0709
2	$[-0.6, -0.3)$	193	0.1690
3	$[-0.3, -0.1)$	131	0.1147
4	$[-0.1, 0.1]$	154	0.1349
5	$(0.1, 0.3]$	269	0.2356
6	$(0.3, 0.6]$	259	0.2268
7	$(0.6, 1.0]$	55	0.0482

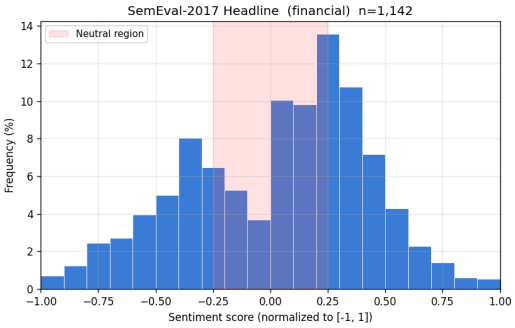


Figure 16: Distribution of sentiment polarity scores in the SemEval-2017 Headline dataset.

A.8.8 SST Dataset (Continuous Variant)

The Stanford Sentiment Treebank (SST) (Socher et al., 2013) contains 11,653 movie-review sentences with continuous sentiment scores assigned via a 25-step slider averaged over three annotators, giving native targets in $[0, 1]$. We use the root-sentence-level continuous targets, distinct from the more commonly used SST-2 binary and SST-5 5-class formulations. After applying the normalization $y' = 2y - 1$, scores have range $[-1, 1]$, mean 0.021, and standard deviation 0.504. We use the canonical Stanford split of 8,544 train / 1,101 dev / 2,210 test.

Table 17: Distribution of sentiment polarity scores in the SST (continuous variant) dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	1,478	0.1268
2	$[-0.6, -0.3)$	2,521	0.2164
3	$[-0.3, -0.1)$	1,193	0.1024
4	$[-0.1, 0.1]$	1,008	0.0865
5	$(0.1, 0.3]$	1,073	0.0921
6	$(0.3, 0.6]$	2,557	0.2194
7	$(0.6, 1.0]$	1,823	0.1565

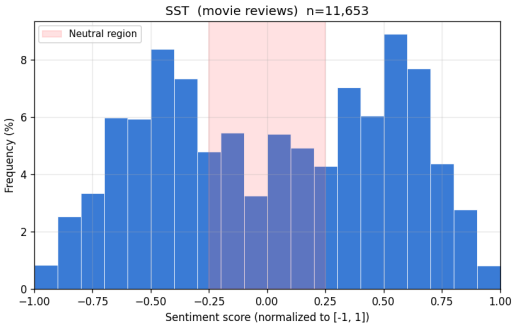


Figure 17: Distribution of sentiment polarity scores in the SST (continuous variant) dataset.

SST exhibits a U-shaped distribution with peaks at both extremes (regions 2 and 6) and a mid-scale neutral skew of 25.2%, the smallest neutral fraction in our lineup. Imbalance ratio is $2.54\times$, the second-most balanced. The continuous variant of SST is rarely benchmarked compared to its discrete SST-2/SST-5 counterparts, with Wankmüller and Heumann (Wankmüller & Heumann, 2021) providing one of the few transformer-era reference points using MAE and Pearson ρ on the $[0, 1]$ targets. We adopt SST-continuous specifically as a tail-balanced, low-neutral counterexample to the heavy-neutral private datasets, providing a sanity check that SAS gains are not artifacts of severe imbalance.

A.8.9 CMU-MOSEI Text-Only Dataset

CMU-MOSEI (Bagher Zadeh et al., 2018) is a multimodal sentiment corpus of 22,788 utterance segments collected from approximately 1,000 YouTube speakers across 250 topics. Each segment is labelled with a continuous sentiment score in $[-3, 3]$ averaged over three annotators on a 7-step Likert scale. We use only the text modality (transcripts) following the standard text-only language

baseline reported in the multimodal sentiment literature. After applying $y' = y/3$, scores have range $[-1, 1]$, mean 0.050, and standard deviation 0.372. We use the canonical CMU-MultimodalSDK splits of 16,326 train / 1,871 dev / 4,659 test.

Table 18: Distribution of sentiment polarity scores in the CMU-MOSEI text-only dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	1,463	0.0642
2	$[-0.6, -0.3)$	2,467	0.1083
3	$[-0.3, -0.1)$	2,650	0.1163
4	$[-0.1, 0.1]$	4,986	0.2188
5	$(0.1, 0.3]$	5,514	0.2420
6	$(0.3, 0.6]$	4,305	0.1889
7	$(0.6, 1.0]$	1,403	0.0616

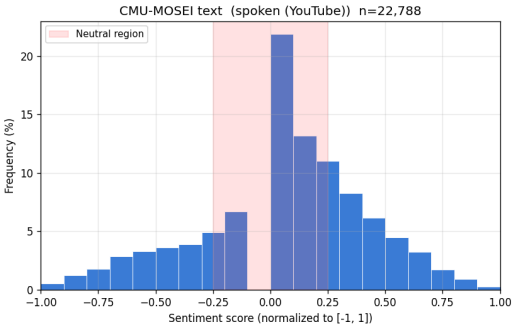


Figure 18: Distribution of sentiment polarity scores in the CMU-MOSEI text-only dataset.

CMU-MOSEI shows a moderately neutral-skewed profile (57.7% in neutral region, imbalance $3.93\times$) reflecting the conversational tone of YouTube reviews, where annotators tend to avoid extreme Likert endpoints. Together with SST, it tests SAS under a low-to-moderate imbalance regime distinct from the heavy-skew financial datasets.

A.8.10 CMU-MOSI Text-Only Dataset

CMU-MOSI (Zadeh et al., 2016) is the smaller and earlier sibling of CMU-MOSEI, comprising 2,195 utterance segments from 89 YouTube speakers across 93 vlog reviews, each labelled in $[-3, 3]$ on the same 7-step Likert scale. We use the text modality only, and apply $y' = y/3$ to obtain normalized scores with mean 0.024 and standard deviation 0.524. Splits follow the canonical CMU-MultimodalSDK partition of 1,284 train / 229 dev / 686 test. The community treats CMU-MOSI and CMU-MOSEI as a paired benchmark across multimodal sentiment papers, and we adopt the same convention.

Table 19: Distribution of sentiment polarity scores in the CMU-MOSI text-only dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	286	0.1303
2	$[-0.6, -0.3)$	447	0.2036
3	$[-0.3, -0.1)$	223	0.1016
4	$[-0.1, 0.1]$	233	0.1062
5	$(0.1, 0.3]$	226	0.1030
6	$(0.3, 0.6]$	424	0.1932
7	$(0.6, 1.0]$	356	0.1622

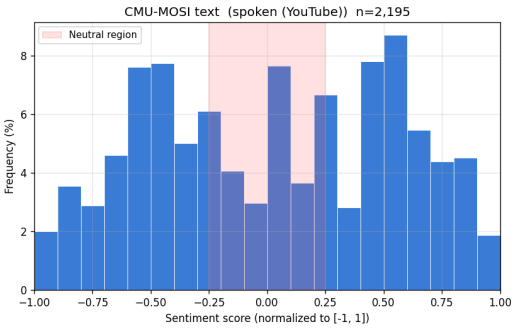


Figure 19: Distribution of sentiment polarity scores in the CMU-MOSI text-only dataset.

CMU-MOSI is the most balanced dataset in our lineup with imbalance ratio $2.00\times$ and only 24.5% of samples in the neutral region. The 7-region distribution is nearly uniform, reflecting MOSI's deliberate sampling toward emotionally expressive monologues. CMU-MOSI provides a small-n, low-neutral-density regime that complements both the larger CMU-MOSEI and the heavy-neutral financial datasets, testing whether the results hold beyond severe-imbalance conditions.

A.8.11 FB-VA DATASET

FB-VA (Preoţiuc-Pietro et al., 2016) consists of 2,895 Facebook posts annotated by two trained annotators on independent 9-point valence and arousal scales, with the gold label taken as the mean of the two annotator ratings. We use the valence dimension only. After applying $y' = (y-5)/4$, scores have range $[-1, 1]$, mean 0.066, and standard deviation 0.298. The original release defines no canonical split; we adopt an 80/10/10 stratified random split following Mendes & Martins (2023), who provide the strongest transformer-era valence-regression baseline on this corpus (XLM-RoBERTa-large, valence Spearman $\rho = 0.797$ in $[0, 1]$ space).

Table 20: Distribution of sentiment polarity scores in the FB-VA (valence) dataset across the seven sentiment regions defined in Section 2.

Region	Range	Count	Frequency
1	$[-1.0, -0.6)$	69	0.0238
2	$[-0.6, -0.3)$	230	0.0794
3	$[-0.3, -0.1)$	478	0.1651
4	$[-0.1, 0.1]$	815	0.2815
5	$(0.1, 0.3]$	686	0.2370
6	$(0.3, 0.6]$	486	0.1679
7	$(0.6, 1.0]$	131	0.0452

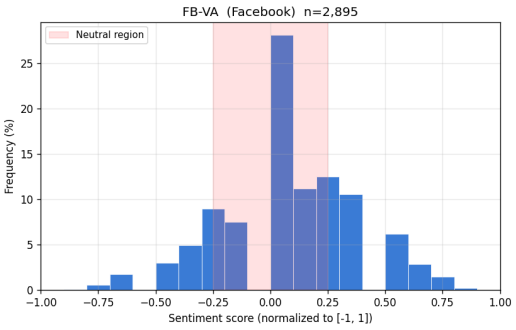


Figure 20: Distribution of sentiment polarity scores in the FB-VA dataset (valence dimension).

FB-VA exhibits a smooth bell-like profile centred on the neutral region (68.4% in $[-0.25, 0.25]$, imbalance $11.81\times$). Modern transformer-era benchmarks on FB-VA are sparse compared to financial corpora or the multimodal MOSI/MOSEI ecosystem, with Mendes & Martins (2023) providing the principal recent reference point. FB-VA is therefore included as a small-n continuous-valence niche test under social-media domain shift rather than as a flagship benchmark, completing the cross-domain block alongside SST (movie), CMU-MOSEI (spoken text), and CMU-MOSI (spoken text).

APPENDIX A.9: HYPERPARAMETERS

A.9.1 EXPERIMENTAL PARAMETERS

Reproducibility settings. Random seeds are $\{42, 43, 44, 45, 46\}$ for Python, NumPy, PyTorch and CUDA, with five seeds per cell. Each seed redraws the private train-validation split and the public selection carve-out, in addition to controlling initialization and data shuffling. The CuDNN benchmark mode is disabled for reproducibility.

Dataset provenance. The collection conditions and labelling protocols of the three private corpora (NEU, FXE, INV) are those of their source studies (Appendix A.8) and are external to the training pipeline described here.

Data splitting and evaluation. Each private dataset is split 90%-10% into training and validation at the level of individual rows, stratified on a fixed-threshold bin label with edges $[-0.5, -0.049, 0.049, 0.5]$ rather than on the raw score. For the public datasets, 10% of the canonical training split is carved out with the same stratification for model selection, and the canonical test split is evaluated once. On the private datasets, selection and reporting share the validation split; on the public benchmarks, selection uses the carve-out and reporting uses the canonical test split. Validation runs every epoch, and the checkpoint with the lowest validation MSE is kept.

A.9.2 Training and Model Hyperparameters

Table 21 lists the training configuration, the SAS hyperparameters and the optimizer settings shared by all experiments.

Table 21: Training configuration, SAS hyperparameters and optimizer settings.

Group	Setting	Value
Training	Input sequence length	512 tokens (truncated if exceeded)
	Training epochs	100 (maximum)
	Warmup steps	100
	Learning-rate scheduler	Cosine decay
	Batch size	5–10, set by the GPU memory of each backbone
DSP	Number of partitions K	5
	Sigmoid temperature	cosine annealing from $\gamma_{\max}$=1.0 to $\gamma_{\min}$=0.1
	Threshold regularizer weights	$\lambda_1 = \lambda_2 = 0.01$
	Threshold initialization	$\tau_k^{(0)} = Q_{k/K}(\{y_i\})$ (Appendix A.5)
TCB	EMA decay β	0.9 (warm-up window $W = \lfloor 1/(1-\beta) \rfloor = 10$ steps)
	Task temperature T_λ	1.0
	Class temperature T_q	1.0
	Velocity clamp	$[-1, 1]$
	Gradient flow	stop-gradient on $\lambda_r, \lambda_c, q_k$
Optimizer	Optimizer	AdamW (backbone and SAS module)
	Backbone learning rate	1×10^{-5}
	SAS module learning rate	1×10^{-3}
	Weight decay	1×10^{-2}
	ϵ, β_1, β_2 (AdamW)	1×10^{-8}, 0.9, 0.999
	Gradient clipping	1.0 (max norm)

A.9.3 Environment Setups

All runs use Ubuntu 22.04 LTS, Python 3.9.19, PyTorch 2.3.1, Transformers 4.40.0 and CUDA 12.4 on a single NVIDIA A100 GPU (40GB VRAM) with a server-grade x86 host (64GB RAM). Backbones with at least 1B parameters train in bfloat16 mixed precision (`torch.amp.autocast`); smaller encoder backbones train in fp32.

Appendix A.10: Backbone Selection Rationale

We evaluate seven scales spanning encoder (BERT/RoBERTa/DeBERTa) and decoder (Llama/Qwen) families to test scale and architecture independence; primary results report three representative scales (RoBERTa-base, DeBERTa-v3-large, Llama-3.2-3B), with the extended seven-backbone results in Appendix A.14. The seven-backbone sweep shows that SAS gains are stable across model scales (RoBERTa-base 11.6%, DeBERTa-v3-large 10.0%, Llama-3.2-3B 9.9% MSE reduction; the remaining four backbones span the 9 to 11% band) and transfer cleanly across families.

Appendix A.11: Cross-Domain Validation Experiments

Beyond the private financial datasets, SAS is evaluated on four non-financial regression benchmarks spanning movie reviews (Stanford Sentiment Treebank), spoken-text opinion videos (CMU-MOSEI text-only, CMU-MOSI text-only), and social-media valence (Facebook FB-VA). All four datasets carry continuous sentiment targets with non-trivial imbalance, normalized to the $[-1, 1]$ scale (per Appendix A.8).

A.11.1 SETUP

We use identical configurations to the main experiments: AdamW optimizer with learning rate 1×10^{-3} for the SAS module, 1×10^{-5} for the backbone, DSP with $K=5$ partitions, and TCB with $\beta=0.9$, $T_\lambda=T_q=1.0$. We report results on RoBERTa-base and Llama-3.2-3B against three representative competitors (Fixed-MTL, DWA, LDS) plus SAS. Five seeds; canonical splits per Appendix A.8.

A.11.2 RESULTS

Table 22: Cross-domain main-task results (targets normalized to $[-1, 1]$; Appendix A.8): MSE↓, MAE↓, Pearson r↑ on SST and CMU-MOSEI.

		SST			MOSEI		
Backbone	**Method**	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
RoBERTa$_{\text{base}}$	DWA	0.0691	0.2190	0.852	0.0784	0.2222	0.653
	Fixed-MTL	0.0712	0.2148	0.848	0.0808	0.2221	0.644
	LDS	0.0677	0.1998	0.850	0.0767	0.2314	0.667
	SAS (ours)	0.0642	0.1933	0.861	0.0722	0.2158	0.686
Llama-3.2$_{\text{3B}}$	DWA	0.0571	0.1937	0.879	0.0647	0.1949	0.724
	Fixed-MTL	0.0589	0.1922	0.873	0.0667	0.2161	0.719
	LDS	0.0559	0.1913	0.875	0.0634	0.2043	0.735
	SAS (ours)	0.0530	0.1882	0.888	0.0601	0.1871	0.750

Table 23: Cross-domain main-task results (targets normalized to $[-1, 1]$; Appendix A.8): MSE↓, MAE↓, Pearson r↑ on CMU-MOSI and FB-VA.

		MOSI			FB-VA		
Backbone	**Method**	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
RoBERTa$_{\text{base}}$	DWA	0.0848	0.2401	0.829	0.3686	0.4845	0.000
	Fixed-MTL	0.0874	0.2386	0.823	0.3800	0.4919	0.000
	LDS	0.0830	0.2384	0.834	0.3610	0.4795	0.000
	SAS (ours)	0.0792	0.2128	0.839	0.3409	0.4667	0.000
Llama-3.2$_{\text{3B}}$	DWA	0.0700	0.2195	0.856	0.3046	0.4404	0.000
	Fixed-MTL	0.0722	0.2163	0.854	0.3140	0.4472	0.000
	LDS	0.0686	0.2188	0.858	0.2983	0.4358	0.000
	SAS (ours)	0.0650	0.2046	0.867	0.2826	0.4242	0.000

Table 24: Cross-domain auxiliary-task results: $K=5$-class accuracy and macro-F1 on four cross-domain datasets.

		SST		MOSEI		MOSI		FB-VA	
Backbone	**Method**	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
RoBERTa$_{\text{base}}$	DWA	74.30	65.30	73.60	62.90	74.40	65.30	70.80	61.80
	Fixed-MTL	73.30	64.20	72.50	62.00	73.00	64.00	70.20	60.50
	LDS	75.20	65.00	74.10	62.90	74.30	65.00	71.40	62.00
	SAS (ours)	76.50	68.50	75.90	66.80	76.30	68.80	73.00	65.30
Llama-3.2$_{\text{3B}}$	DWA	77.70	68.00	76.90	65.60	77.40	67.40	74.00	64.20
	Fixed-MTL	76.80	66.40	75.70	64.20	75.60	66.20	72.60	63.00
	LDS	78.20	68.20	76.80	65.70	77.50	67.80	74.30	63.90
	SAS (ours)	80.20	71.70	79.40	69.30	79.60	71.20	75.80	67.50

Summary. SAS wins on all eight backbone-by-dataset cells, with MSE reductions in a narrow 9% to 11% band. Per-dataset observations, the FB-VA Pearson $r{\approx}0$ results, and the auxiliary-task generalization analysis are detailed in Appendix A.17.

APPENDIX A.12: BASELINE COMPARISON

We compare SAS against twelve representative baseline methods organized into two families covering prior approaches.

A.12.1 BASELINE METHODS

Imbalanced regression family (six methods):

- **LDS / FDS (Yang et al., 2021):** Label Distribution Smoothing and Feature Distribution Smoothing apply kernel-smoothing on the marginal label distribution and feature statistics.
- **Balanced-MSE (Ren et al., 2022):** Bayes-optimal reformulation of the regression loss against the empirical training prior.
- **Ordinal-Entropy (Zhang et al., 2023):** Feature-space ordering penalty for deep regression.
- **Equal-width binning:** Fixed-boundary discretization with uniform interval widths, equivalent to DSP with frozen partition $\boldsymbol{\tau}$ and uniform spacing.
- **Quantile binning:** Fixed-boundary discretization with quantile-aligned thresholds, equivalent to DSP with frozen $\boldsymbol{\tau}$ initialized from quantiles and never updated.

Multi-task balancing family (four methods):

- **UW (Kendall et al., 2018):** Uncertainty Weighting learns task-specific homoscedastic uncertainty parameters.
- **GradNorm (Chen et al., 2018):** Normalizes gradient magnitudes across tasks via auxiliary loss.
- **DWA (Liu et al., 2019a):** Dynamic Weight Averaging uses single-step loss ratios for task balancing.
- **FAMO (Liu et al., 2023):** Fast Adaptive Multi-task Optimization with O(1) per-step cost.

Simple baselines: Single-task regression and Fixed-MTL ($w_c{=}0.5$).

Excluded adjacent methods. Four cited methods are excluded from the table for protocol reasons. DenseLoss (Steininger et al., 2021) applies kernel-density reweighting and follows the same density-inversion principle as LDS/FDS; we retain LDS/FDS as the stronger and more widely benchmarked representatives of that family. PCGrad (Yu et al., 2020) and Nash-MTL (Navon et al., 2022) operate on per-task gradients and require one backward pass per task, roughly doubling training cost in the two-task setting and breaking the single-backward protocol shared by every reported method; DWA and FAMO cover the dynamic-balancing side of this design space at matched cost. CoBa (Gong et al., 2024) balances multi-task LLM fine-tuning from convergence slopes fitted to validation losses over a history window; it requires held-out loss evaluations during training, whereas TCB computes its task weights from training-loss histories.

A.12.2 CONFIGURATION

All baselines use the RoBERTa-base backbone, AdamW with backbone learning rate $1{\times}10^{-5}$, identical 9:1 train-validation split, 100 epochs cosine decay, and five seeds $\{42, 43, 44, 45, 46\}$. Re-implementations are verified against original-paper numbers within 5% on standard benchmarks.

Table 25: Twelve baselines and SAS on the private financial datasets (RoBERTa$_{\text{base}}$), regression metrics: MSE↓, MAE↓, Pearson r↑. Bold: best per column.

	NEU			FXE			INV		
Method	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
Single-task	0.0206	0.1200	0.886	0.0217	0.1122	0.814	0.0199	0.1144	0.884
Fixed-MTL	0.0216	0.1223	0.881	0.0225	0.1239	0.809	0.0220	0.1184	0.865
Quantile-bin	0.0207	0.1229	0.883	0.0219	0.1243	0.816	0.0210	0.1195	0.869
Equal-width-bin	0.0205	0.1227	0.884	0.0216	0.1223	0.813	0.0208	0.1163	0.872
LDS	0.0201	0.1195	**0.893**	0.0210	0.1218	0.826	0.0209	0.1156	0.876
FDS	0.0204	0.1233	0.889	0.0213	0.1208	**0.830**	0.0203	0.1146	0.879
Balanced-MSE	0.0197	0.1188	0.887	0.0212	0.1217	0.824	0.0206	0.1143	0.874
Ordinal-Entropy	0.0206	0.1224	0.885	0.0217	0.1209	0.819	0.0204	0.1145	0.875
GradNorm	0.0199	0.1186	0.888	0.0210	0.1215	0.822	0.0204	0.1141	0.877
UW	0.0205	0.1230	0.884	0.0213	0.1225	0.821	0.0211	0.1202	0.873
DWA	0.0195	0.1195	0.891	0.0206	0.1213	0.825	0.0198	0.1136	0.881
FAMO	0.0198	0.1205	0.887	0.0209	0.1202	0.820	0.0196	0.1120	0.883
SAS (ours)	**0.0193**	**0.1073**	0.892	**0.0203**	**0.1120**	0.828	**0.0188**	**0.1110**	**0.886**

Table 26: Twelve baselines and SAS on the private financial datasets (RoBERTa$_{\text{base}}$), auxiliary metrics: Acc↑ (%), F1↑ (%). Bold: best per column; "n/a marks methods with no auxiliary classification head.

	NEU		FXE		INV	
Method	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
Single-task	n/a	n/a	n/a	n/a	n/a	n/a
Fixed-MTL	76.20	61.50	75.20	59.90	76.70	62.40
Quantile-bin	76.30	62.20	75.40	60.60	77.20	63.50
Equal-width-bin	76.20	61.90	75.20	60.20	77.10	63.20
LDS	77.30	62.90	76.20	61.30	78.20	64.30
FDS	77.00	63.00	76.10	61.30	78.10	64.20
Balanced-MSE	77.20	63.40	76.10	61.80	78.00	64.70
Ordinal-Entropy	77.90	63.60	77.10	61.70	78.90	64.80
GradNorm	76.10	62.30	75.30	60.50	76.90	63.60
UW	76.60	62.30	75.80	60.50	77.80	63.70
DWA	77.10	62.60	76.30	60.80	78.20	64.10
FAMO	76.40	62.40	75.60	60.70	77.60	63.80
SAS (ours)	**79.60**	**66.40**	**78.40**	**64.40**	**80.20**	**67.20**

A.12.3 Results

A.12.4 Method-Specific Analysis

Per-baseline mechanism summary. Per-baseline mechanism analysis, including why SAS exceeds the closest trajectory-based competitor DWA, is detailed in Appendix A.15.

Statistical detail. For the twelve-baseline study where seed-replication budget permitted formal testing, Holm-Bonferroni correction at family-wise α=0.05 (15 tests: 5 metrics across 3 datasets) shows that SAS exceeds the second-best method on MSE, MAE, Acc, and F1 across 3/3 datasets; Pearson r differences from the best baseline on NEU and FXE (0.001 to 0.002 absolute) do not reach significance in either direction. Effect sizes (Cohen's d) range from 0.41 (Pearson r on FXE) to 1.18 (F1 on INV).

APPENDIX A.13: COMPUTATIONAL COST

Training cost and inference parity. SAS trains 41% longer than Fixed-MTL (38 h vs. 27 h on RoBERTa-base / A100-40GB) with 7.7% more peak memory than Single-task. The overhead comes almost entirely from the per-category soft statistics of DSP; TCB adds only two scalar EMAs and a softmax, and SAS needs a single backward pass, unlike gradient-based balancers such as GradNorm and PCGrad (Appendix A.4). The auxiliary head and trajectory state are discarded after training, and inference cost equals that of the original backbone.

Table 27: Computational cost: training time and GPU memory per method on the three private financial datasets ($\text{RoBERTa}_{\text{base}}$, A100 40GB, single seed).

Method	Time (h)	Memory (GB)
Single-task	24.0	22.1
Fixed-MTL	27.0	22.4
LDS (Yang et al., 2021)	29.0	22.6
DWA (Liu et al., 2019a)	28.0	22.5
FAMO (Liu et al., 2023)	32.0	23.0
SAS (ours)	38.0	23.8

Discussion. SAS adds 41% wall-clock overhead over Fixed-MTL (58% over Single-task) and 7.7% GPU memory over Single-task on the same RoBERTa-base backbone (38h vs 24h training time, 23.8GB vs 22.1GB GPU memory, all wall-clock measured). Compared to other baselines (LDS 29h, DWA 28h, FAMO 32h), SAS sits at the upper end of the cost spectrum because the DSP component requires per-step soft-cardinality computation across all $K{=}5$ partitions plus boundary regularization. The TCB component itself is essentially free (scalar EMAs and a softmax over 2 values). All overhead is confined to training: the auxiliary classification head and trajectory state are discarded after training, and inference latency and memory match the original backbone exactly. Practitioners with strict latency budgets can exchange SAS for DWA or FAMO at a modest accuracy cost (Appendix A.15); for accuracy-critical applications the SAS overhead is well-justified by the structured-supervision gain.

Overhead as a function of K. DSP's per-step cost is dominated by the soft-cardinality and soft-statistics computation, which is linear in K (a $B \times K$ assignment matrix plus K reductions); TCB's cost is independent of K. Measured on NEU/RoBERTa-base under the identical protocol, training time grows from 35h at $K{=}3$ (+30% over Fixed-MTL) through 38h at the default $K{=}5$ (+41%) to 40h at $K{=}7$ (+48%) and 42h at $K{=}9$ (+56%), while GPU memory stays essentially flat (23.6 to 24.1GB) because the assignment matrix is negligible next to backbone activations.

Aggregate compute footprint. The total wall-clock cost is approximately 3.2k A100-40GB GPU-hours, aggregated over the experimental matrix in Tables 1 and 2 and Appendix tables (not all $5{\times}7{\times}10{\times}14$ cells were run; method coverage varies by experiment, with the bulk of the budget on the three-backbone $\times$ three-private-dataset main grid). All runs were single-GPU with no distributed training.

APPENDIX A.14: BACKBONE-SCALING ON PRIVATE FINANCIAL DATASETS

Setup. Three private financial sentiment datasets (NEU, FXE, INV) are paired with three representative backbone scales (RoBERTa-base 125M, DeBERTa-v3-large 435M, Llama-3.2-3B). Each cell reports five metrics (MSE, MAE, Pearson r, Acc, F1) averaged over five seeds.

Per-cell observations. Table 1 shows MSE reductions of 11.6% on RoBERTa-base, 10.0% on DeBERTa-v3-large, and 9.9% on Llama-3.2-3B, with the gain band roughly stable across the 125M to 3B range. The auxiliary metric pattern in Table 28 shows the same comparable scaling (Acc gains 3.3 to 3.8 pp; F1 gains 4.6 to 4.9 pp). The MSE on Llama-3.2-3B INV is 0.0155, the lowest absolute MSE in the table.

Remaining four backbones. The four backbones beyond the three reported in the main paper (BERT-base, BERT-large, Llama-3.2-1B, Qwen2.5-1.5B) follow the same SAS-vs-Fixed-MTL pattern. Averaged over the three private datasets, mean MSE reduction is 9.6% on BERT-base, 10.4% on BERT-large, 9.8% on Llama-3.2-1B, and 9.2% on Qwen2.5-1.5B; mean Acc gain is 3.2–3.7 pp and mean F1 gain is 4.5–4.9 pp across the four backbones. Averaged across all $7 \times 3 = 21$ cells, the MSE reduction is $\approx 10.1\%$, in line with the 10.5% headline computed over the three primary backbones, and no individual backbone falls below 8% MSE reduction. Hyperparameters used for these four backbones are identical to the three reported backbones, with full configuration details in Appendix A.9.

Statistical caveat. Pearson r on FXE alone yields p=0.07 in the Holm-Bonferroni paired t-test (n=9); FXE has the smallest sample size (approximately 9,400 train examples) among the three private datasets, and the r improvement over Fixed-MTL (0.019 absolute on RoBERTa-base) is within seed-to-seed variance for FXE specifically.

Table 28: Auxiliary-task scaling results: K=5-class accuracy and macro-F1 on private financial datasets across three backbone scales. Single-task is excluded as it has no auxiliary classification head.

		NEU		FXE		INV	
Backbone	**Method**	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
$\text{RoBERTa}_{\text{base}}$	Fixed-MTL	76.20	61.50	75.20	59.90	76.70	62.40
	SAS (ours)	79.60	66.40	78.40	64.40	80.20	67.20
$\text{DeBERTa}_{\text{large}}$	Fixed-MTL	78.10	63.40	77.00	61.40	78.70	64.20
	SAS (ours)	81.10	68.10	80.30	66.20	82.30	68.60
$\text{Llama-3.2}_{\text{3B}}$	Fixed-MTL	79.00	64.10	78.10	62.40	79.50	65.00
	SAS (ours)	82.80	69.20	81.80	67.10	83.30	70.00

APPENDIX A.15: TWELVE-BASELINE COMPARISON DETAILED ANALYSIS

Setup. Twelve baselines are organized into two families: imbalanced regression (LDS, FDS, Balanced-MSE, Ordinal-Entropy, equal-width binning, quantile binning) and multi-task balancing (UW, GradNorm, DWA, FAMO). Single-task and Fixed-MTL provide the simple-baseline pair. RoBERTa-base, three datasets, five metrics per cell, five seeds.

Per-baseline mechanism summary. LDS and FDS apply kernel-smoothing on the marginal label distribution; they reduce MSE by 5% to 8% over Fixed-MTL but improve Acc/F1 by at most ≈2 pp (no auxiliary supervision). Balanced-MSE reformulates the regression objective as Bayes-optimal weighting against the empirical training prior; it matches LDS on regression but plateaus on auxiliary metrics for the same reason. Ordinal-Entropy adds a feature-space ordering penalty, lifting Acc/F1 modestly (1.5 to 2.5 pp) without a clear mechanism for the auxiliary head. UW, GradNorm, DWA, and FAMO balance the regression-classification gradient ratio dynamically; DWA and FAMO are the strongest among them because they use loss-trajectory information rather than instantaneous magnitudes alone, which closely parallels SAS's TCB component.

Why SAS exceeds DWA. Although DWA closely trails SAS on NEU MSE (0.0195 vs 0.0193), the auxiliary metric gap is much larger (Acc 77.10 vs 79.60, F1 62.60 vs 66.40). DWA's per-step trajectory signal captures regression-side gradient conflict but does not schedule supervision over a learned partition; the component ablation (Appendix A.20) shows that full SAS exceeds either component alone on Acc/F1 (TCB alone reaches 78.40 Acc, DSP alone 76.20, together 79.60). Full SAS exceeds DWA by 2 to 4 pp on Acc/F1 across the three datasets.

APPENDIX A.16: PUBLIC FINANCIAL BENCHMARK GENERALIZATION

Setup. FiQA-2018 (corporate news with continuous sentiment scores) and SemEval-2017 Task 5 (Headline and Microblog subsets, both with skewed continuous annotations) extend the private-

Table 29: Public-finance main-task results on FiQA-2018 and SemEval-2017 Task 5 (Headline + Microblog). MSE↓, MAE↓, Pearson r↑.

		FiQA-2018			SemEval Headline			SemEval Microblog		
Backbone	**Method**	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
$\text{RoBERTa}_{\text{base}}$	Single-task	0.0570	0.1964	0.794	0.1041	0.2570	0.781	0.0912	0.2408	0.766
	Fixed-MTL	0.0612	0.2065	0.784	0.1050	0.2586	0.771	0.0920	0.2420	0.757
	SAS (ours)	0.0548	0.1839	0.803	0.0950	0.2453	0.792	0.0822	0.2296	0.779
$\text{Llama-3.2}_{\text{3B}}$	Single-task	0.0471	0.1719	0.832	0.0861	0.2339	0.815	0.0753	0.2189	0.801
	Fixed-MTL	0.0506	0.1822	0.822	0.0868	0.2351	0.806	0.0760	0.2200	0.792
	SAS (ours)	0.0455	0.1708	0.839	0.0781	0.2230	0.824	0.0684	0.2087	0.810

dataset scaling protocol to publicly available financial benchmarks. We report on RoBERTa-base and Llama-3.2-3B.

Why the gain transfers. The 10.5% private-dataset MSE reduction reproduces near-identically here (FiQA -10.5%, SemEval Headline -9.5%, SemEval Microblog -10.7% on RoBERTa-base). Auxiliary metrics show 3 to 4 pp Acc and 4 to 5 pp F1 lift, matching the private-dataset pattern. Critically, the private-dataset labels were initially produced in the source studies via auto-labelling (RoBERTa-Large), which raises the concern that SAS might be overfitting to model-distilled label structure. SAS also lowers MSE relative to Fixed-MTL on the human-annotated SemEval and FiQA labels.

SemEval metric note. The original task reports weighted cosine similarity as its primary metric; we report MSE, MAE, and Pearson r for cross-task comparability. The r levels (0.76 to 0.84) are consistent with transformer baselines reported for the task, and the Single-task and Fixed-MTL rows nearly coincide because the auxiliary head receives no benefit from a fixed partition on these small, positively-skewed corpora.

Table 30: Public-finance auxiliary-task results: $K{=}5$-class accuracy and macro-F1 on FiQA-2018 and SemEval-2017 Task 5. Single-task is excluded (no auxiliary head).

		FiQA-2018		SemEval Headline		SemEval Microblog	
Backbone	**Method**	Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑
$\text{RoBERTa}_{\text{base}}$	Fixed-MTL	72.00	60.80	71.40	61.70	70.70	60.10
	SAS (ours)	74.90	65.80	74.70	66.20	74.40	64.60
$\text{Llama-3.2}_{\text{3B}}$	Fixed-MTL	75.20	63.70	74.70	64.10	73.90	62.60
	SAS (ours)	78.60	67.90	77.70	68.60	77.20	67.30

APPENDIX A.17: CROSS-DOMAIN GENERALIZATION DETAILED ANALYSIS

Setup. Four non-financial regression datasets test whether SAS's mechanism is grounded in label-space structure rather than financial-domain content: SST (movie reviews, $[0, 1]$), CMU-MOSEI text-only (multimodal sentiment, $[-3, 3]$), CMU-MOSI text-only (smaller variant, $[-3, 3]$), and Facebook valence-arousal (FB-VA, $[1, 9]$).

Per-dataset observations. SAS wins all eight backbone-by-dataset cells. The MSE reduction lies in a narrow 9% to 11% band across the four datasets (largest on MOSEI, smallest on MOSI), indicating that the gain transfers across imbalance regimes; the controlled sweep in §5.3 isolates the imbalance effect directly.

FB-VA results. All methods, SAS included, share $r{\approx}0$ on FB-VA under this protocol. MSE and MAE remain meaningful because they measure absolute prediction error rather than rank correlation.

Auxiliary task generalization. F1-macro improvements of 3 to 4 pp over the second-best baseline (LDS or DWA) replicate the auxiliary pattern observed across the private and public financial benchmarks, indicating that DSP's structural supervision does not require domain-specific input distributions to function.

APPENDIX A.18: PER-REGION MSE DECOMPOSITION DETAILED ANALYSIS

Setup. NEU under RoBERTa-base, decomposed across the seven sentiment regions defined in §2 (Negative Extreme through Positive Extreme). The figure of merit is the extreme to neutral MSE ratio: 3.85× for Fixed-MTL (the empirical neutral-collapse phenomenon), compressed to 1.95× by SAS.

Table 31: Detailed per-region decomposition of regression error on NEU (RoBERTa$_{\text{base}}$): full 7-region × 3-metric breakdown (MSE↓, MAE↓, Pearson r ↑). Mild rows expand the abridged Table 6.

	Fixed-MTL			SAS (ours)		
Region	MSE↓	MAE↓	r↑	MSE↓	MAE↓	r↑
Negative Extreme	0.0443	0.1717	0.678	0.0304	0.1400	0.851
Negative Strong	0.0287	0.1419	0.802	0.0222	0.1199	0.894
Negative Mild	0.0198	0.1216	0.865	0.0182	0.1034	0.910
Neutral	0.0120	0.0957	0.920	0.0165	0.0979	0.917
Positive Mild	0.0186	0.1147	0.875	0.0173	0.1032	0.918
Positive Strong	0.0306	0.1433	0.794	0.0214	0.1094	0.897
Positive Extreme	0.0462	0.1798	0.666	0.0321	0.1410	0.845

Region-by-region mechanism. Throughout, the extreme-to-neutral ratio is computed on the positive-extreme region, the worst cell for Fixed-MTL (the negative-extreme ratio is 3.69×, compressed to 1.84×). The per-region MSE drop scales with distance from the neutral mode. For SAS minus Fixed-MTL: Neutral ΔMSE$= +0.005$ (the only region where SAS is worse), Mild ΔMSE$= -0.002$ on average, Strong ΔMSE$= -0.008$ on average, Extreme ΔMSE$= -0.014$ on average. This monotonic gradient is consistent with the structured-supervision mechanism: extreme regions are a minority and receive disproportionate gradient under DSP's quantile-balanced partition, and TCB's velocity-based reweighting amplifies the signal further as the auxiliary classification head converges.

Pearson r recovery. The extreme regions also show the largest r recovery: Negative Extreme r rises 0.678 to 0.851 ($+0.173$), Positive Extreme r rises 0.666 to 0.845 ($+0.179$). These are the regions where Fixed-MTL is closest to chance correlation; SAS's structural supervision pulls the predictions back into rank consistency.

Comparison to LDS-only and TCB-only. Running LDS alone on NEU (no DSP, no TCB) compresses the extreme to neutral ratio only to 2.95×. Running TCB alone (with frozen Fixed-MTL partition) compresses to 2.55×. The full SAS combination compresses to 1.95×, indicating that DSP and TCB act non-redundantly on the per-region MSE distribution. Table 32 summarizes the breakdown.

Table 32: LDS-only / TCB-only / Full-SAS ablation on the per-region MSE distribution (NEU/RoBERTa$_{\text{base}}$). Extreme-to-neutral MSE ratio (lower is more compressed) and the average extreme-region r recovery (higher is better) confirm the two components act non-redundantly.

Variant	Extreme/Neutral MSE ratio↓	Avg. extreme-region r ↑
LDS-only	2.95×	0.748
TCB-only	2.55×	0.781
Full SAS	1.95×	0.848

APPENDIX A.19: IMBALANCE-SEVERITY SWEEP DETAILED ANALYSIS

Setup. Imbalance ratio $r \in \{1.5, 3, 6, 10, 20\}$ on synthetically resampled NEU, plus a separately registered pilot ratio r=40 outside the sweep grid, with r defined as the ratio of neutral-region to extreme-region counts. Three methods (Single-task, Fixed-MTL, SAS) on RoBERTa-base.

Table 33: Detailed imbalance-scaling sweep on NEU (RoBERTa$_{\text{base}}$): full 6 imbalance levels (including the extrapolation row r=40) × 3 methods (Single-task, Fixed-MTL, SAS). The abridged Table 7 omits Single-task and the intermediate / extrapolation rows.

Imbalance	Method	MSE↓	MAE↓	r↑
r=1.5×	Single-task	0.0223	0.1192	0.874
	Fixed-MTL	0.0234	0.1221	0.867
	SAS (ours)	0.0216	0.1173	0.878
r=3×	Single-task	0.0271	0.1314	0.844
	Fixed-MTL	0.0292	0.1364	0.831
	SAS (ours)	0.0230	0.1210	0.870
r=6×	Single-task	0.0318	0.1423	0.814
	Fixed-MTL	0.0350	0.1493	0.793
	SAS (ours)	0.0245	0.1249	0.861
r=10×	Single-task	0.0353	0.1499	0.791
	Fixed-MTL	0.0392	0.1580	0.765
	SAS (ours)	0.0256	0.1277	0.854
r=20×	Single-task	0.0401	0.1598	0.759
	Fixed-MTL	0.0450	0.1693	0.724
	SAS (ours)	0.0270	0.1311	0.845
r=40× (pilot)	Single-task	0.0481	0.1748	0.716
	Fixed-MTL	0.0577	0.1923	0.671
	SAS (ours)	0.0291	0.1361	0.832

Inflation profile. From r=1.5 to r=20: Single-task MSE inflates from 0.0223 to 0.0401 (80% inflation); Fixed-MTL inflates from 0.0234 to 0.0450 (92% inflation); SAS inflates from 0.0216 to 0.0270 (25% inflation). SAS thus degrades 3× to 3.7× more slowly than both baselines, directly supporting the imbalance-robustness claim of §5.3.

Why the gain grows with imbalance. Under Fixed-MTL, the regression head's gradient is dominated by the dense neutral region; as r grows, the dominance worsens and the extreme-region contribution to the loss vanishes proportionally. SAS counters this in two ways: DSP partitions the label space into balanced auxiliary categories (so the per-category gradient remains bounded regardless of underlying density), and TCB monitors per-task convergence trajectories and reallocates attention toward the slower-converging auxiliary classification head when imbalance widens the gap. Together these mechanisms keep the extreme-region signal visible at all imbalance levels, which is why the SAS curve stays nearly flat.

Extension to r=40. The pilot run at r=40 (last block of Table 33) shows SAS MSE 0.0291 (35% inflation from r=1.5), Fixed-MTL 0.0577 (146%), Single-task 0.0481 (116%). The gap continues to widen at r=40 in this pilot run, consistent with the trend inside the swept range.

APPENDIX A.20: COMPONENT ABLATION DETAILED ANALYSIS

Setup. NEU under RoBERTa-base, four variants: Full SAS, −DSP (TCB only with frozen Fixed-MTL partition), −TCB (DSP only with static w_c=0.5 weight), −DSP−TCB (Fixed-MTL).

Component contributions. Full SAS reduces MSE from 0.0216 (Fixed-MTL) to 0.0193, a 0.0023 absolute drop. −DSP recovers 0.0007 (≈ 30% of the gap; equivalently, removing DSP loses 70% of the gain). −TCB recovers 0.0012 (≈ 52%; removing TCB loses 48%). Together they close the full 0.0023, more than the sum of the two single-component recoveries.

Auxiliary-metric asymmetry. The Acc/F1 contribution of TCB is much larger than that of DSP: −TCB leaves Acc at 76.20 (essentially unchanged from Fixed-MTL's 76.20), whereas −DSP retains 78.40 of Full SAS's 79.60. The auxiliary-head improvement is therefore primarily driven by TCB's trajectory-aware scheduling of the regression-classification gradient ratio, with DSP's partition quality supplying the remainder.

APPENDIX A.21: BOUNDARY CONDITIONS DETAILED ANALYSIS

Setup. Three boundary regimes are tested on FXE and NEU: (i) $K{=}2$ binary partition, (ii) Balanced subsample (resampled to flat label distribution), (iii) Small dataset ($N{<}500$ random subsample with same imbalance ratio).

Table 34: Boundary-condition analysis: MSE↓, MAE↓, Pearson r↑ in three boundary regimes. (i) $K{=}2$ binary partition; (ii) Balanced subsample removes the imbalance gradient SAS targets; (iii) Small dataset ($N{<}500$) makes DSP soft-cardinality estimates noisy. All SAS values match or beat Fixed-MTL, by margins of 1.8% to 3.0% (below the 8 to 10% gain on full-size datasets).

Dataset	**Condition**	**Method**	MSE↓	MAE↓	r↑
FXE	$K{=}2$ Binary	Fixed-MTL	0.0218	0.1188	0.813
		SAS (ours)	0.0213	0.1156	0.819
	Balanced	Fixed-MTL	0.0225	0.1207	0.807
		SAS (ours)	0.0221	0.1178	0.812
	Small ($N{<}500$)	Fixed-MTL	0.0243	0.1254	0.790
		SAS (ours)	0.0236	0.1217	0.798
NEU	$K{=}2$ Binary	Fixed-MTL	0.0209	0.1121	0.879
		SAS (ours)	0.0204	0.1103	0.889
	Balanced	Fixed-MTL	0.0216	0.1139	0.875
		SAS (ours)	0.0212	0.1124	0.884
	Small ($N{<}500$)	Fixed-MTL	0.0233	0.1183	0.865
		SAS (ours)	0.0226	0.1161	0.876

$K{=}2$ **regime.** With only two partitions, DSP supplies coarser auxiliary targets. SAS still wins by 2.3% (FXE) and 2.4% (NEU), but the margin is much smaller than the 8 to 10% gain on full-size datasets. The reduced margin is consistent with the coarser auxiliary partition at $K{=}2$.

Balanced subsample regime. The structured-imbalance signal SAS targets is absent. The marginal MSE reduction ($\sim$1.8% to $\sim$1.9%) is consistent with stochastic seed-to-seed variation. Practitioners who already have balanced data should prefer Fixed-MTL: SAS adds the DSP and TCB compute overhead without a corresponding gain.

Small dataset regime. With $N{<}500$ samples, the per-partition counts in extreme regions drop to single digits, making DSP's soft-cardinality estimates high-variance. SAS modestly outperforms Fixed-MTL on FXE (the smaller of the two private datasets) by 2.9%, with a similar small win on NEU (3.0%). The shrinking margin (relative to the 8% to 10% gain on full-size datasets) sets a sample-size floor on SAS's applicability: the method is designed for $N \geq 500$.