LOSS-GUIDED TSETLIN MACHINES: MAKING THE LEARNING OBJECTIVE EXPLICIT

Anonymous authors
Paper under double-blind review

ABSTRACT

The Tsetlin Machine (TM) has no explicit notion of a loss function. Its learning signal is fixed to an implicit, margin-based update rule (MBUR), with no way to choose an objective function suited to a specific task. This paper introduces Loss-Guided TM, which replaces MBUR with an explicit and swappable loss function. In our proposed method, clause feedback and weight updates are driven by the loss, while keeping the Tsetlin Automaton and clause logic unchanged. The approach incorporates cross entropy (CE), symmetric cross entropy (SCE), and asymmetric loss (ASL). We evaluate Loss-Guided TM across six datasets that span multiclass, multilabel, and binary classification tasks. For each dataset, the best-performing Loss-Guided configuration is significantly better calibrated than the MBUR-based one, reducing calibration error by a factor of 3 to 11. It further improves accuracy and F1 across all datasets. The gains are greatest for the imbalanced multilabel datasets, where ASL improves macro F1 by up to 9 percentage points compared to MBUR. A qualitative inspection shows that the clause-based global interpretability, examined at the per-class level on image and text data, is preserved when using the Loss-Guided mechanism. These results show that our explicit TM learning objective allows direct targeting of uncertainty calibration, data imbalance, and task structure, opening the TM to the broader literature on loss functions.

1 INTRODUCTION

The Tsetlin Machine (TM) is a logic-based alternative to neural networks, applied across various domains, such as image classification (Granmo et al., 2019; Grønningsæter et al., 2024), text categorization and natural language understanding (Berge et al., 2019; Saha et al., 2022; 2023; Bhattarai et al., 2021; Gao et al., 2026), regression (Darshana Abeyrathna et al., 2019), contextual bandits (Seraj et al., 2022), and board-game winner prediction (Giri et al., 2022). Its predictions are disjunctions of conjunctive clauses over binarized input literals, directly readable as propositional logic. Together with a novel learning strategy (finding common denominators and reasoning by elimination), TMs become easier to interpret than neural networks (Yadav et al., 2022; Shende et al., 2025; Tarasyuk et al., 2026). The same bit-level structure makes it efficient to run, with reported reductions in computation cost and memory footprint over comparable neural architectures (Kundu et al., 2026). This combination of competitive accuracy, interpretability, and efficiency positions the TM as a practical, transparent alternative to deep learning (Lei et al., 2020).

A loss function defines what a model is actually being trained to do. Different loss functions can produce different learned behavior even from the same architecture and the same data (Wasserman, 2004). Cross entropy, the standard choice for classification, corresponds to the model's output representing a categorical or Bernoulli distribution over the labels, while mean squared error corresponds to a Gaussian distribution (Goodfellow et al., 2016). More broadly, loss functions in the literature are designed around the structure of a specific problem. This includes, among many others (Wang et al., 2022), losses built for extreme class imbalance (Lin et al., 2020), for maximizing the margin between classes (Cortes & Vapnik, 1995), for region overlap in segmentation (Milletari et al., 2016), for matching entire distributions rather than individual predictions (Arjovsky et al., 2017), and for balancing reconstruction accuracy against a regularized latent representation (Kingma & Welling, 2014).

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, TMs do not have an explicit loss function as a neural network does. Instead, a TM uses a heuristic update probability (Eq. 2), which is the normalized absolute error between the model's vote sum and the voting threshold (T). This update mechanism, here termed margin-based update rule (MBUR), is hardwired directly into the TM's learning algorithm, with no principled bridge to the concept of loss functions. The TM therefore cannot optimize its performance using any of the well-established objective functions, tailored to the various learning tasks. This paper introduces a Loss-Guided learning mechanism for the TM that replaces MBUR's heuristic update probability with an explicit loss function, while retaining the Tsetlin Automaton based logical rule learning. We evaluate this mechanism using three different loss functions, cross entropy (CE), symmetric cross entropy (SCE), and asymmetric loss (ASL), assessing both predictive performance and interpretability against MBUR across six datasets.

2 BACKGROUND

TM is a machine learning model built from finite-state learning automata and propositional logic (Granmo, 2018). The Tsetlin Automaton (TA) is the model's basic learning unit, a finite-state machine that learns to select one of two actions, include or exclude, based on reward and penalty signals from its environment. Each input feature is binarized and paired with its negation, forming the set of literals. Each literal is assigned a TA, which decides whether to include or exclude the literal. Thus, a single propositional logic clause is defined by a dedicated team of TAs that operate on the literal set. In brief, the clause computes the conjunction $c_k = \bigwedge_{f \in L_k} f$ of the literals its TAs choose to include, where L_k is the subset of literals included by clause k's automata. A clause is active, $c_k = 1$, when every literal in L_k is true for the current input, and inactive, $c_k = 0$, otherwise.

The TM is initialized with the hyperparameters C clauses, threshold T, specificity s, and negative-class sampling rate q. Here, we are considering the weighted coalesced variant of the TM (Glimsdal & Granmo, 2021). Thus, each of the C clauses is shared among the classes. Each clause has a weight per class, w_{ik} for clause k and class i, whose sign gives the clause's polarity for that class. Weights are initialized uniformly at random from $\{-1, 1\}$ and move in discrete ± 1 steps during training (Abeyrathna et al., 2021). The class sum (v_i) is the total vote of the model towards class i and is calculated as:

$$v_i = \sum_k w_{ik}\, c_k. \tag{1}$$

For multiclass classification, the predicted class is the one with the highest class sum. For multilabel classification, a class is predicted whenever its class sum is positive.

During training, the probability of updating a clause (P_i) from feedback is given by:

$$P_i = \frac{|y_i \cdot T - \text{clip}(v_i, -T, T)|}{2T}, \qquad y_i \in \{-1, 1\}. \tag{2}$$

When there are multiple classes, the probability of a class with label 0 giving feedback is further gated by probability $q/(M-1)$, where M is the number of classes. Thus, q controls the amount of negative feedback given by classes with label 0 and is one of the key parameters in multilabel classification. A clause can receive three types of feedback, namely, ***Recognize*** (Type-Ia), ***Erase*** (Type-Ib), and ***Reject*** (Type-II). For a class i with label 1, the active positive polarity clauses receive Type-Ia, the inactive positive polarity clauses receive Type-Ib, and the active negative polarity clauses receive Type-II feedback. Similarly, for a class i with label 0, the active positive polarity clauses receive Type-II feedback, inactive negative polarity clauses receive Type-Ib, and active negative polarity clauses receive Type-Ia feedback. Each feedback type and how the literals are updated in each case is explained in detail in Granmo (2018). In this paper, we refer to the TM trained under this margin-based update rule simply as MBUR.

3 METHODOLOGY

Given a sample x and its labels y, each clause k is evaluated against x, resulting in a clause output $c_k \in \{0, 1\}$. In the proposed loss-guided scheme, the model output $\hat{y}_i$ for the class i is given by,

$$\hat{y}_i = A(z_i), \qquad z_i = \left(\frac{v_i}{C/2}\right) \tag{3}$$

where A is an activation function, v_i is the vote sum (Eq. 1), and z_i is the normalized vote sum. The normalization uses $C/2$ because the clauses are divided into sets of positive and negative polarity clauses, each with opposite learning objectives. In the experiments, A is the softmax function for multiclass classification and sigmoid otherwise. Let $\ell(y, \hat{y})$ be the loss function. The loss $L = \ell(y, \hat{y})$ then quantifies the discrepancy between the predictions $\hat{y}$ of the model and the labels y.

Clause updates We estimate the effect a clause has on the loss (L) by calculating ΔL_k.

$$\Delta L_k = L - L_{-k}, \qquad L_{-k} = \ell(y, \hat{y}_{-k}) \tag{4}$$

where $\hat{y}_{-k}$ is the model output obtained by inverting the clause output to $\bar{c}_k$, and L_{-k} is the loss using this new output. Thus, ΔL_k quantifies whether inverting clause k's output reduces the loss. We then evaluate the probability of updating a clause k, P_k, as

$$P_k = 1 - e^{-\lambda|\Delta L_k|} \tag{5}$$

where $\lambda > 0$ is the feedback rate parameter. The function $1 - e^{-\lambda x}$ is the cumulative distribution function of the exponential distribution with rate λ, mapping the non-negative correction signal $|\Delta L_k|$ to a valid probability. Unlike MBUR, where each class's vote sum is used to calculate update probability, here each clause only receives one update regardless of the number of classes. The type of feedback a clause receives is determined by ΔL_k and c_k, as shown in Eq. 6.

$$\text{feedback}(c_k, \Delta L_k) = \begin{cases} \text{Type Ia} & \text{if } c_k = 1 \text{ and } \Delta L_k < 0 \\ \text{Type Ib} & \text{if } c_k = 0 \text{ and } \Delta L_k > 0 \\ \text{Type II} & \text{if } c_k = 1 \text{ and } \Delta L_k > 0 \\ \text{None} & \text{otherwise} \end{cases} \tag{6}$$

Weight Updates Clause weights $w_{ik} \in \mathbb{R}$ are initialized uniformly at random in $[-1, 1]$. The weights are updated deterministically, and are decoupled from the stochastic TA feedback scheme. The weight update follows the gradient of the loss with respect to each weight, obtained via the chain rule through $\hat{y}_i = A(z_i)$:

$$\nabla w_{ik} = \frac{\partial \ell}{\partial w_{ik}} = \frac{\partial \ell}{\partial \hat{y}_i} \cdot \frac{\partial \hat{y}_i}{\partial z_i} \cdot \frac{\partial z_i}{\partial w_{ik}} = \ell'(y_i, \hat{y}_i) \cdot A'(z_i) \cdot \frac{c_k}{C/2} \tag{7}$$

Let $g_i = \ell'(y_i, \hat{y}_i) \cdot A'(z_i) = \partial \ell / \partial z_i$ denote the combined loss-and-activation gradient with respect to the normalized vote. For all the ℓ and A pairs used in the experiments here, g_i has a closed form. Absorbing the constant factor $C/2$ into the learning rate η gives the weight update,

$$w_{ik} \leftarrow w_{ik} - \eta \, g_i \, c_k. \tag{8}$$

Hyperparameter changes The proposed mechanism does not depend on the *threshold* (T). Similarly, the hyperparameter q is also no longer needed, since in the proposed approach, a clause receives feedback at most once per sample. Instead, these are replaced by the *learning rate* (η) and the *feedback rate* (λ).

Algorithm 1 summarizes one training epoch under the proposed mechanism.

4 Experimental Design

4.1 Datasets

MNIST (LeCun et al., 1998) is a standard image classification benchmark, containing 60,000 training and 10,000 test grayscale images across 10 digit classes. **FashionMNIST** (Xiao et al., 2017) is similar to MNIST, with the same train/test split sizes across 10 clothing-item classes. **CelebA** (Liu et al., 2015) is a multilabel face-attribute classification dataset, containing 162,770 training and 19,867 test images, each labeled with 40 binary facial attributes. This is a highly imbalanced dataset and contains 9.03 labels per sample on average. For this dataset, images are resized to 64×64 RGB and random horizontal flipping is applied during training. Since the TM needs binarized inputs, we

Algorithm 1 One training epoch

Require: training set $\{(x_n, y_n)\}_{n=1}^{N}$, learning rate η, feedback rate λ

for each training sample (x, y) **do**
  for each clause $k = 1, \ldots, C$ **do**
    Evaluate clause k against x, giving clause output $c_k \in \{0, 1\}$
  end for
  for each class i **do**
    $z_i \leftarrow \frac{2}{C} \sum_k w_{ik}\, c_k$
    $\hat{y}_i \leftarrow A(z_i)$ ▷ Eq. 3
  end for
  $L \leftarrow \ell(y, \hat{y})$
  for each clause $k = 1, \ldots, C$ **do**
    for each class i **do**
      $z_i' \leftarrow z_i - \frac{2}{C}(c_k - \bar{c}_k)\, w_{ik}$ ▷ normalized vote with inverted clause output $\bar{c}_k$
      $\hat{y}_i' \leftarrow A(z_i')$
    end for
    $\Delta L_k \leftarrow L - \ell(y, \hat{y}')$ ▷ Eq. 4
    $P_k \leftarrow 1 - e^{-\lambda |\Delta L_k|}$ ▷ Eq. 5
    if $u \sim \text{Uniform}(0, 1) < P_k$ **then**
      Apply feedback$(c_k, \Delta L_k)$ to clause k ▷ Eq. 6
    end if
    for each class $i = 1, \ldots, M$ **do**
      $g_i \leftarrow \nabla_{z_i} L$
      $w_{ik} \leftarrow w_{ik} - \eta\, g_i\, c_k$ ▷ Eq. 8
    end for
  end for
end for

use 8-level thermometer encoding (Buckman et al., 2018) on the pixel values, except for MNIST, which is thresholded using a static pixel intensity of 75.

IMDB is a binary movie review sentiment classification task, containing 25,000 training and 25,000 test examples. The text is preprocessed by removing stopwords and encoding words into binary bag-of-words (unigram and bigram) features. We then select the top 12,000 features using a chi-squared test. **AG News** (Zhang et al., 2015) is a multiclass text classification task, containing 120,000 training and 7,600 test examples across 4 news-topic classes. We follow preprocessing similar to IMDB, with an additional lemmatization step, using only unigram features, and selecting the top 2,000 features. **GoEmotions** (Demszky et al., 2020) is a multilabel emotion classification task, containing 43,400 training and 5,400 test examples across 28 emotion labels. This is a highly imbalanced dataset, averaging 1.18 labels per sample. Text is tokenized with a tweet-aware tokenizer that preserves case, collapses repeated characters, and properly handles emojis and unicode characters. We then remove stopwords, lemmatize, replace URLs with a placeholder, and encode the top 10,000 features selected from unigram and bigram bag-of-words.

4.2 Implementation and Evaluation Details

MBUR and Loss-Guided TM are compared under matched capacity, meaning the clauses and specificity s hyperparameters are kept the same across the two models. All the models are trained for 100 epochs and use coalesced clauses. We did not conduct an extensive hyperparameter grid search to optimize each condition. Instead, hyperparameters for MBUR were selected through initial trials based on prior literature and understanding of T and q. For Loss-Guided TM, a similar approach was used to select the learning rate η and feedback rate λ. Table 1 lists the selected hyperparameters used for each experiment.

We use three different loss functions across datasets to demonstrate the use of different losses according to the dataset. Cross entropy (CE) is the standard classification loss, measuring the divergence between the predicted and true class distributions. Symmetric cross entropy (SCE) (Wang et al.,

2019) adds a reverse cross entropy term that stays bounded even when the model is confidently wrong, improving robustness to label noise. Asymmetric loss (ASL) (Ridnik et al., 2021) applies separate down-weighting factors to positive and negative predictions along with a hard threshold on the gradient for confident negatives. This is designed to counteract the positive-negative imbalance inherent to multilabel classification. For brevity, we refer to the Loss-Guided TM trained with CE, SCE, or ASL simply by the name of the loss.

Table 1: Hyperparameters used for each dataset and condition. T and q are used by MBUR, while η (learning rate) and λ (feedback rate) are used by Loss-Guided TM.

Dataset	C	s	T	q	$p \times p$	ℓ	η	λ
MNIST	1000	10.0	2000	1.0	10×10	CE	1.0	1.0
FashionMNIST	16000	10.0	20000	1.0	3×3	CE	0.5	3.0
IMDB	10000	1.1	10000	1.0	-	CE	0.01	1.0
AG News	8000	2.0	40000	1.0	-	CE SCE (α=1.0, β=1.0)	0.01 0.005	1.0 1.0
GoEmotions	22400	2.0	25000	1.5	-	CE ASL (γ_+=0.0, γ_-=4.0, clip=0.01)	0.5 0.05	4.0 6.0
CelebA	80000	30.0	80000	10.0	3×3	CE ASL (γ_+=2.0, γ_-=4.0, clip=0.01)	1.0 1.0	25.0 25.0

For multiclass classification tasks, accuracy, macro F1, macro ROC-AUC, macro PR-AUC, and macro ECE (expected calibration error) at the final epoch are reported. ECE measures how well a model's predicted confidence matches how often it is actually correct. It is relevant whenever a confidence score is used for more than just picking the top prediction, for example flagging uncertain cases for review.

For multilabel classification, macro F1, weighted F1, ROC-AUC, PR-AUC, and macro ECE are reported instead. Accuracy is uninformative for multilabel classification and is therefore omitted. ROC-AUC and PR-AUC are threshold-independent ranking metrics, better suited to imbalanced and multilabel settings than a fixed-threshold metric like F1. PR-AUC is the more informative of the two under heavy label imbalance, since ROC-AUC can stay high even when a classifier performs poorly on the rare positive label.

Each dataset experiment is repeated with 10 seeds for each of the conditions. Differences in the metrics are tested with the paired Wilcoxon signed-rank test (Wilcoxon, 1945). Each dataset involves multiple pairwise comparisons. Thus, raw p-values are adjusted with the Holm-Bonferroni correction (Holm, 1979) to control the family-wise error rate. Effect size is measured alongside every p-value as paired Cohen's d_z (Lakens, 2013), to estimate the magnitude of the difference in the results. In the tables 2 and 3, † denotes significance ($p < 0.05$) relative to MBUR.

5 Results

Table 2 shows the test set metrics for the multiclass classification datasets. Across all the datasets tested here, the Loss-Guided TM achieves substantially better calibration, with a large and consistent effect size. Accuracy and F1 are comparable or better than MBUR. The largest gain in accuracy was seen on the IMDB dataset, with a very large effect size ($d_z \approx 18.7$), reflecting highly consistent and low-variance improvement across seeds. On AG News, SCE achieves the highest accuracy of all conditions, at the cost of a smaller calibration gain compared to CE.

Table 3 shows the results for multilabel classification datasets. ASL achieves significantly better F1 scores (macro and weighted) on both datasets, with large effect sizes. However, ASL also produces the least calibrated models, showing calibration error higher than MBUR. Similar to the multiclass results, the CE condition achieved the lowest calibration error on both datasets. For CelebA, the CE condition also achieved slightly better PR and ROC AUC scores, alongside weighted F1 comparable to ASL. However, ASL achieves substantially higher macro F1. This reflects the design of ASL, which counteracts multilabel imbalance and helps the model learn rare classes.

Table 2: Multiclass classification test metrics (mean ± std over 10 seeds).

Dataset	Condition	Accuracy (%)	F1 (%)	ROC-AUC (%)	PR-AUC (%)	ECE
MNIST	MBUR	**98.75 ± 0.11**	**98.74 ± 0.11**	99.92 ± 0.01	99.46 ± 0.06	0.0056 ± 0.0002
	CE	98.55 ± 0.13	98.54 ± 0.13	**99.98 ± 0.00**†	**99.87 ± 0.02**†	**0.0016 ± 0.0001**†
FashionMNIST	MBUR	90.32 ± 0.13	90.28 ± 0.15	98.85 ± 0.04	92.49 ± 0.07	0.0263 ± 0.0009
	CE	**90.49 ± 0.18**†	**90.46 ± 0.16**†	**99.31 ± 0.01**†	**95.37 ± 0.08**†	**0.0047 ± 0.0003**†
IMDB	MBUR	85.52 ± 0.12	85.52 ± 0.12	90.20 ± 0.10	87.37 ± 0.23	0.0934 ± 0.0011
	CE	**88.61 ± 0.08**†	**88.61 ± 0.08**†	**95.14 ± 0.04**†	**94.96 ± 0.04**†	**0.0142 ± 0.0012**†
AG News	MBUR	88.94 ± 0.24	88.91 ± 0.24	96.60 ± 0.09	90.48 ± 0.26	0.0648 ± 0.0039
	CE	89.74 ± 0.15†	89.73 ± 0.15†	**97.71 ± 0.02**†	**94.32 ± 0.05**†	**0.0109 ± 0.0006**†
	SCE	**90.38 ± 0.13**†	**90.36 ± 0.13**†	97.50 ± 0.02†	93.98 ± 0.06†	0.0265 ± 0.0006†

Table 3: Multilabel classification results on the test set (mean ± std over 10 seeds).

Dataset	Condition	F1 (macro, %)	F1 (weighted, %)	PR-AUC (%)	ROC-AUC (%)	ECE
GoEmotions	MBUR	28.09 ± 1.08	38.03 ± 0.63	28.51 ± 0.60	80.55 ± 0.28	0.1601 ± 0.0076
	CE	26.38 ± 0.37	40.42 ± 0.25†	30.69 ± 0.22†	81.60 ± 0.16†	**0.0142 ± 0.0005**†
	ASL	**37.06 ± 0.28**†	**50.44 ± 0.18**†	**34.46 ± 0.15**†	**82.45 ± 0.08**†	0.2396 ± 0.0008
CelebA	MBUR	55.21 ± 0.67	67.99 ± 0.51	57.82 ± 0.16	86.08 ± 0.18	0.1457 ± 0.0150
	CE	58.47 ± 0.50†	71.12 ± 0.25†	**67.04 ± 0.10**†	**89.60 ± 0.04**†	**0.0299 ± 0.0021**†
	ASL	**63.86 ± 0.11**†	**73.75 ± 0.09**†	66.73 ± 0.08†	89.44 ± 0.05†	0.1802 ± 0.0009

Averaged over the ten rarest attributes, ASL's mean F1 gain over MBUR is +15.8 percentage points, compared to only +5.8 points over the ten most common attributes, a negative correlation between attribute frequency and gain, while CE shows no such correlation. The effect is clearest on the rarest attributes. On Bald (2.3% of samples), ASL raises F1 from $30.0 \pm 8.6\%$ to $57.4 \pm 0.7\%$, and on Blurry (5.1% of samples), from $14.0 \pm 4.4\%$ to $38.3 \pm 1.1\%$, in both cases also sharply reducing the across-seed variance. CE does not share this pattern. On Mustache (4.1% of samples), CE lowers F1 relative to MBUR ($30.6\% \rightarrow 27.6\%$), while ASL still improves it (42.4%). On the most common attributes the gap narrows sharply, e.g. No_Beard (83.4% of samples) improves only from $91.5 \pm 0.4\%$ to $94.4 \pm 0.1\%$.

On GoEmotions, classes grief (0.18% of samples) and relief (0.35% of samples) score close to zero F1 across all three conditions. Where MBUR does show a non-zero mean, for example $5.1 \pm 10.8\%$ on grief, the variance is quite large. This reflects the stochastic nature of TM learning, and indicates that this class in particular is not reliably learned. ASL's F1 gain over MBUR grows with class frequency instead of shrinking, the opposite direction from CelebA. On more frequent classes this gain is large, for example on amusement (5.4% of samples), F1 rises from $47.4 \pm 3.8\%$ to $77.5 \pm 0.7\%$, and on gratitude (6.1% of samples), from $66.0 \pm 4.4\%$ to $86.8 \pm 1.1\%$. On some of the rarer classes ASL falls behind MBUR instead, for example on embarrassment (0.70% of samples), F1 falls from $25.2 \pm 4.3\%$ to $7.7 \pm 2.6\%$. The CE condition had the worst macro F1 of the three, since it was particularly difficult for it to learn some of the rarer classes. On disapproval (4.7% of samples), CE lowers F1 from $17.1 \pm 1.2\%$ to $10.2 \pm 1.5\%$, while ASL raises it to $23.1 \pm 1.2\%$.

Figure 1 shows the train and test curves over 100 training epochs for accuracy (for multiclass) or macro F1 (for multilabel) metrics. The Loss-Guided models show a steadier increase and more stable learning curves compared to MBUR. MBUR overfits heavily on the IMDB dataset, where the train accuracy reaches 100% within the first few epochs. In contrast, the Loss-Guided model shows a slower learning curve and lower degree of overfitting, while achieving higher test accuracy. On AG News, GoEmotions, and CelebA, MBUR's performance peaks within the first few epochs, followed by a decline in both training and test performance. In contrast, Loss-Guided TM improves over more epochs, reaching higher peak test performance before showing signs of overfitting.

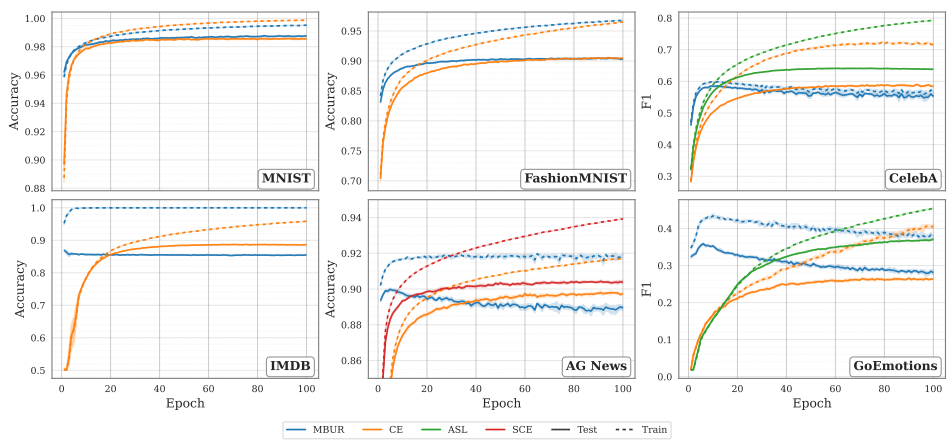


Figure 1: Training and test curves over 100 epochs for all six datasets. Plots show accuracy for MNIST, FashionMNIST, IMDB, and AG News, and macro F1 for CelebA and GoEmotions. Solid curves denote test performance and dashed curves denote training performance for MBUR and Loss-Guided variants.

5.1 Interpretability

Beyond the predictive metrics reported above, we examine the global interpretations produced by the models. Local interpretability highlights the parts of the input responsible for a given prediction, computed using the WAC score (Clarke et al., 2026), while global interpretability aggregates all the learned patterns for a particular class. Global interpretability therefore allows us to analyze and compare the combined contribution of all the clauses in the model. We present this comparison for the CelebA and IMDB datasets. Due to the stochastic nature of TM learning, each seed can produce a different set of clauses and slightly different interpretations. Thus, we compare the interpretations averaged over the 10 seeds.

For CelebA, we follow the global interpretability procedure for convolutional TM described by Shende et al. (2025). Figure 2 shows these global interpretations for the 40 CelebA attributes. Non-gray regions indicate pixels the model considers important, and gray regions indicate don't-care or unimportant regions. CE and ASL produce visually similar interpretations across all 40 attributes. For several attributes, MBUR already focuses on the correct patterns with reasonably strong importance, for example Bushy_Eyebrows, Wearing_Lipstick, and Smiling. These are typically the classes for which MBUR has fairly high F1 score as well. However, it fails to capture the correct pattern for several other classes. On Wearing_Earrings, MBUR shows no class-specific patterns. On Wearing_Necktie, MBUR focuses on the face rather than the relevant region. On Blond_Hair, Wearing_Hat, and Wearing_Necklace, MBUR gives very low importance to the relevant regions. In all these cases, the Loss-Guided models produce better interpretations, and clearly capture more patterns for the classes. MBUR also fails to distinguish between related attributes that Loss-Guided separates correctly. For instance, MBUR shows nearly the same pattern for Mustache, Sideburns, and Goatee classes, while Loss-Guided models capture distinct patterns for the three attributes. The same holds for Big_Nose and Chubby, where MBUR's interpretations are close to identical, while CE and ASL produce clearly different patterns for the two. More generally, Loss-Guided's interpretations cover a noticeably larger part of the face, whereas MBUR remains sparse and concentrated on a few small regions.

For IMDB, we calculate the interpretability using a procedure similar to that for the image dataset. However, here we additionally include the negative polarity clauses in the final importance score. For each class, we take the weighted sum of positive-polarity clauses, which gives the combined contribution for the positive-polarity clauses. Since the clauses contain non-negated and negated literals, we take a difference between the two parts to produce an effective per-feature contribution. We repeat the same procedure for negative-polarity clauses to obtain their per-feature contribution. The difference of the two polarity contributions, followed by the asymmetric normalization procedure of Shende et al. (2025) gives the combined per-feature contribution of all the clauses for a class. These



Figure 2: Per-class global interpretations for all 40 CelebA facial attributes, comparing MBUR, Loss-Guided CE, and Loss-Guided ASL averaged over 10 seeds.

are visualized as word clouds per class in Figure 3. Each class's own topic vocabulary emerges as positive evidence (shown here as red), and other classes' vocabulary as negative evidence (shown here as blue), consistently across the conditions.

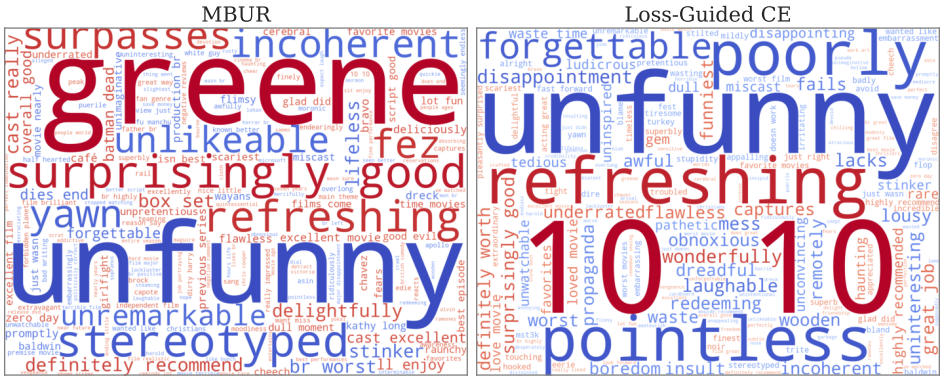


Figure 3: Per-class global interpretation for the IMDB dataset under MBUR and Loss-Guided CE. Words in red indicate positive word importance, while blue indicates negative importance. The relative magnitude of the importance is shown using the size of the word and the intensity of the color.

Generic movie-review vocabulary with no inherent sentiment receives near-zero importance under both conditions, for example "story" (0.0033 vs 0.0024), "style" (0.0033 vs 0.0043), and "watching" (-0.0048 vs 0.0052), given as Loss-Guided vs MBUR. A few words also receive near-zero importance despite sounding sentiment-laden, but these are simply rare in the training data. "Stupid ending" and "absolutely horrendous" appear in only 6 and 7 training reviews respectively. These occurrences are too few for either model to learn a reliable pattern from. The two conditions differ most on the words attributed with a high importance. Loss-Guided gives canonical sentiment words

a higher importance than MBUR, for both polarities, for example "worst" (-0.71 vs -0.35), "awful" (-0.70 vs -0.24), "10 10" (1.00 vs 0.66), and "wonderfully" (0.84 vs 0.49). Loss-Guided also gives sentiment-irrelevant proper nouns a lower importance than MBUR, for example "greene" (0.46 vs 1.00), "fez" (0.28 vs 0.90), "wayans" (-0.43 vs -0.74), and "kathy long" (-0.30 vs -0.76).

6 Discussion

Loss-Guided TM's advantage over MBUR is not confined to a single metric. It holds across performance, calibration, and training stability, with the choice of loss determining where the largest gain falls. Loss-Guided matches or improves accuracy and F1 over MBUR on every dataset tested, and is consistently better calibrated. The amount of improvement is determined by the loss chosen. CE gives the best calibration but is not designed for imbalanced, multilabel tasks, while ASL trades calibration for performance in that setting. This shows different losses are better suited for different tasks.

Because the Loss-Guided mechanism updates the TAs using the established feedback mechanism, the model still learns Boolean propositional rules over the features, preserving interpretability by construction. The classes for which Loss-Guided models perform better, also produces better interpretations, as seen for the CelebA dataset. MBUR interpretations are weakest on the rare attributes where its F1 is lowest, whereas Loss-Guided TM concentrates its largest F1 gains on those same rare classes, producing clearer interpretations.

Loss-Guided learning is also more stable than MBUR (Figure 1). MBUR's early-peak-then-decline pattern on GoEmotions, CelebA, and AG News is different from overfitting, since train and test decline together, indicating degradation of the model itself. In our limited tuning for the MBUR model, the same pattern was observed with several tested configurations, mostly in the case of multilabel datasets. Loss-Guided models showed no such decline on any of the datasets, staying stable or continuing to improve for the rest of training.

Hyperparameters used in the experiment were fixed to one operating point per dataset, with fixed shared hyperparameters between MBUR and every Loss-Guided variant to isolate the effect of the loss-guided learning mechanism. These results describe a common point of comparison, not the best achievable configuration for either mechanism. A proper hyperparameter search should be done in future studies to identify optimal configurations and systematically evaluate the remaining hyperparameters.

Making the TM's objective explicit opens access to a much larger design space than the three losses tested here. Loss functions are a decades-deep, mature area of machine learning research, developed for regression, imbalance, robustness, and structured prediction, and MBUR's fixed feedback rule gave the TM no way to draw on any of it. The same substitution mechanism should also extend beyond a single loss term to regularizers, penalties added to shape what a model learns rather than just how it is scored. How far this substitution mechanism extends, and what it makes possible beyond the classification tasks tested here, needs further research.

7 Conclusion

This paper presented a replacement of MBUR, the Tsetlin Machine's fixed, implicit feedback rule, with an explicit, swappable loss function, keeping the underlying Tsetlin Automata and clause logic unchanged. Across six datasets spanning multiclass, multilabel, and binary classification, Loss-Guided TM either matches or improves F1 and accuracy over MBUR, and achieves better calibration in nearly every case. Furthermore, specific losses align with specific task structures, as demonstrated by ASL achieving substantially higher performance on imbalanced multilabel datasets. The Loss-Guided mechanism does not adversely affect the clause-based interpretability of the TM, and on some datasets produces clearer interpretations than MBUR. By making the learning objective explicit, Loss-Guided TM opens the Tsetlin Machine to the broader loss function literature, developed over decades for calibration, imbalance, and task structure, giving a direct way to target these properties in future work.

AI USE STATEMENT

We used generative AI tools for tasks with recommended disclosure: drafting and editing sections of this paper, improving readability and correcting grammar, creating or modifying scientific figures, help with latex formatting.

We have not used generative AI tools for any of the tasks requiring disclosure under the ICLR 2027 AI Policy for Authors: generating synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results.

We have reviewed all AI-assisted work: every drafted or edited passage was checked against the underlying experiments and source literature before inclusion. 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 Background and Methodology sections give the full specification of both the MBUR and Loss-Guided mechanisms. Table 1 lists every hyperparameter used per dataset and condition. Dataset preprocessing is described in detail in the Datasets section. All experiments use 10 fixed seed values, shared across every condition, so the paired comparisons and reported results can be reproduced exactly. All code implementing the method, the experiments, and the full experimental results will be made publicly available upon acceptance.

AUTHOR CONTRIBUTIONS

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

ACKNOWLEDGMENTS

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

REFERENCES

K. Darshana Abeyrathna, Ole-Christoffer Granmo, and Morten Goodwin. Extending the Tsetlin Machine with Integer-Weighted Clauses for Increased Interpretability. *IEEE Access*, 9:8233–8248, 2021. doi: 10.1109/ACCESS.2021.3049569.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 214–223, 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

Geir Thore Berge, Ole-Christoffer Granmo, Tor Oddbj ørn Tveit, Morten Goodwin, Lei Jiao, and Bernt Viggo Matheussen. Using the Tsetlin Machine to Learn Human-Interpretable Rules for High-Accuracy Text Categorization With Medical Applications. *IEEE Access*, 7:115134–115146, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2935416.

Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Measuring the Novelty of Natural Language Text Using the Conjunctive Clauses of a Tsetlin Machine Text Classifier. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021)*, pp. 410–417. SciTePress, 2021. doi: 10.5220/0010382204100417.

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer Encoding: One Hot Way To Resist Adversarial Examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=S18Su--CW.

Paul F. A. Clarke, Ivan Cmelo, Runar Helin, Mayur Kishor Shende, Ole-Christoffer Granmo, and Darren Fayne. The Tsetlin Machine: A "Third Way" in QSAR Modeling. *Journal of Chemical Information and Modeling*, May 2026. ISSN 1549-9596. doi: 10.1021/acs.jcim.5c03109.

Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

K. Darshana Abeyrathna, Ole-Christoffer Granmo, Xuan Zhang, Lei Jiao, and Morten Goodwin. The regression Tsetlin machine: a novel approach to interpretable nonlinear regression. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 378(2164):20190165, 2019. ISSN 1364-503X. doi: 10.1098/rsta.2019.0165.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. doi: 10.18653/v1/2020.acl-main.372.

Jiechao Gao, Rohan Kumar Yadav, Yuangang Li, Yuandong Pan, Jie Wang, Ying Liu, and Michael Lepech. LLM-Guided Semantic Bootstrapping for Interpretable Text Classification with Tsetlin Machines. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30210–30222, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.1510. URL https://aclanthology.org/2026.findings-acl.1510/.

Charul Giri, Ole-Christoffer Granmo, Herke Van Hoof, and Christian D. Blakely. Logic-based AI for Interpretable Board Game Winner Prediction with Tsetlin Machine. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9, 2022. doi: 10.1109/IJCNN55064.2022.9892796.

Sondre Glimsdal and Ole-Christoffer Granmo. Coalesced Multi-Output Tsetlin Machines with Clause Sharing. *arXiv preprint arXiv:2108.07594*, 2021. doi: 10.48550/arXiv.2108.07594.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN 978-0-262-03561-3. URL https://www.deeplearningbook.org/.

Ole-Christoffer Granmo. The Tsetlin Machine – A Game Theoretic Bandit Driven Approach to Optimal Pattern Recognition with Propositional Logic. *arXiv preprint arXiv:1804.01508*, 2018. doi: 10.48550/arXiv.1804.01508.

Ole-Christoffer Granmo, Sondre Glimsdal, Lei Jiao, Morten Goodwin, Christian W. Omlin, and Geir Thore Berge. The Convolutional Tsetlin Machine, 2019. URL https://arxiv.org/abs/1905.09688.

Ylva Grønningsæter, Halvor S. Smørvik, and Ole-Christoffer Granmo. An Optimized Toolbox for Advanced Image Processing with Tsetlin Machine Composites. In *2024 International Symposium on the Tsetlin Machine (ISTM)*, pp. 1–8, 2024. doi: 10.1109/ISTM62799.2024.10931429.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 1979. doi: 10.2307/4615733.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014. doi: 10.48550/arXiv.1312.6114.

Souraja Kundu, Shruti S. Patkar, Saras Mani Mishra, Gaurav Trivedi, and Farhad Merchant. A Comprehensive Review of Tsetlin Machines: Concepts, Applications, Analysis, and the Future. *IEEE Internet of Things Journal*, 13(10):20105–20127, 2026. ISSN 2327-4662. doi: 10.1109/JIOT.2026.3665145.

Daniël Lakens. Calculating and Reporting Effect Sizes to Facilitate Cumulative Science: A Practical Primer for t-Tests and ANOVAs. *Frontiers in Psychology*, 2013. doi: 10.3389/fpsyg.2013.00863.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Jie Lei, Adrian Wheeldon, Rishad Shafik, Alex Yakovlev, and Ole-Christoffer Granmo. From Arithmetic to Logic based AI: A Comparative Analysis of Neural Networks and Tsetlin Machine. In *2020 27th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, pp. 1–4, 2020. doi: 10.1109/ICECS49266.2020.9294877.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2018.2858826.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. doi: 10.1109/ICCV.2015.425.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 565–571, 2016. doi: 10.1109/3DV.2016.79.

Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric Loss for Multi-Label Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. doi: 10.1109/ICCV48922.2021.00015.

Rupsa Saha, Ole-Christoffer Granmo, Vladimir I. Zadorozhny, and Morten Goodwin. A relational tsetlin machine with applications to natural language understanding. *Journal of Intelligent Information Systems*, 59(1):121–148, 2022. ISSN 1573-7675. doi: 10.1007/s10844-021-00682-5.

Rupsa Saha, Ole-Christoffer Granmo, and Morten Goodwin. Using Tsetlin Machine to discover interpretable rules in natural language processing applications. *Expert Systems*, 40(4):e12873, 2023. ISSN 1468-0394. doi: 10.1111/exsy.12873.

Raihan Seraj, Jivitesh Sharma, and Ole-Christoffer Granmo. Tsetlin Machine for Solving Contextual Bandit Problems. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=b-WnRS7kSEN`.

Mayur Kishor Shende, Ole-Christoffer Granmo, Runar Helin, Vladimir I. Zadorozhny, and Rishad Shafik. A Methodology for Transparent Logic-Based Classification Using a Multi-Task Convolutional Tsetlin Machine. In *2025 International Symposium on the Tsetlin Machine (ISTM)*, pp. 9–16, October 2025. doi: 10.1109/ISTM67926.2025.00010.

Olga Tarasyuk, Anatoliy Gorbenko, Tousif Rahman, Lei Jiao, Ole-Christoffer Granmo, Rishad Shafik, and Alex Yakovlev. Learning dynamics, pattern recognition capability and interpretability of the Tsetlin Machine. *Pattern Recognition*, 174:113028, 2026. ISSN 0031-3203. doi: 10.1016/j.patcog.2025.113028.

Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, 9:187–212, 2022. doi: 10.1007/s40745-020-00253-5.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric Cross Entropy for Robust Learning With Noisy Labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. doi: 10.1109/ICCV.2019.00041.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. doi: 10.1007/978-0-387-21736-9.

Frank Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1945. doi: 10.2307/3001968.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 2017. doi: 10.48550/arXiv.1708.07747.

Rohan Kumar Yadav, Lei Jiao, Ole-Christoffer Granmo, and Morten Goodwin. Robust interpretable text classification against spurious correlations using and-rules with negation. In Lud De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4439–4446. International Joint Conferences on Artificial Intelligence Organization, 7 2022. doi: 10.24963/ijcai.2022/616. URL https://doi.org/10.24963/ijcai.2022/616. Main Track.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level Convolutional Networks for Text Classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/250cf8b51c773f3f8dc8b4be867a9a02-Paper.pdf.

A Appendix

A.1 MNIST and FashionMNIST

For MNIST and FashionMNIST, we compare the global interpretation for both conditions, following the same convolutional TM procedure of Shende et al. (2025) used for CelebA. We additionally include the local interpretation for a few random samples from the test set. Figures 4 and 5 show the local and global interpretations for the two datasets, averaged over 10 seeds. MBUR and CE are visually indistinguishable in both.

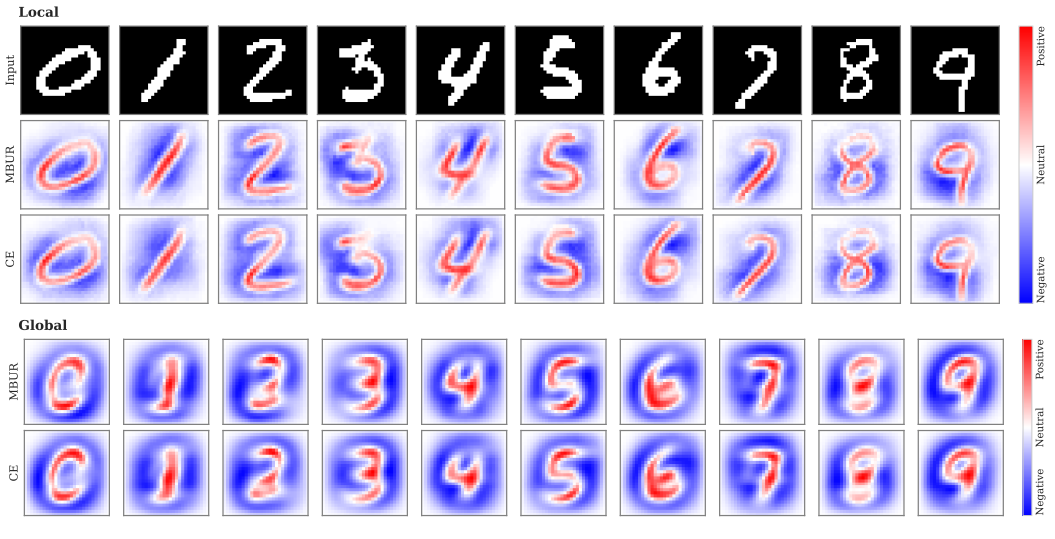


Figure 4: The local and global interpretation for the MNIST dataset. Red indicates positive feature importance, while blue indicates negative importance. The color intensity indicates the relative magnitude of the importance. The white region (≈ zero importance) indicates *don't care* features, which do not affect the predictions of the model.

A.2 IMDB

IMDB showed the largest gap between MBUR and Loss-Guided, both in accuracy and in overfitting behavior. Before settling on the reported configuration, we swept clause count from 500 to 20,000, with the threshold T and specificity s scaled accordingly, across twelve trial configurations (Table 4). MBUR overfit under every evaluated configuration, with train accuracy reaching 94%

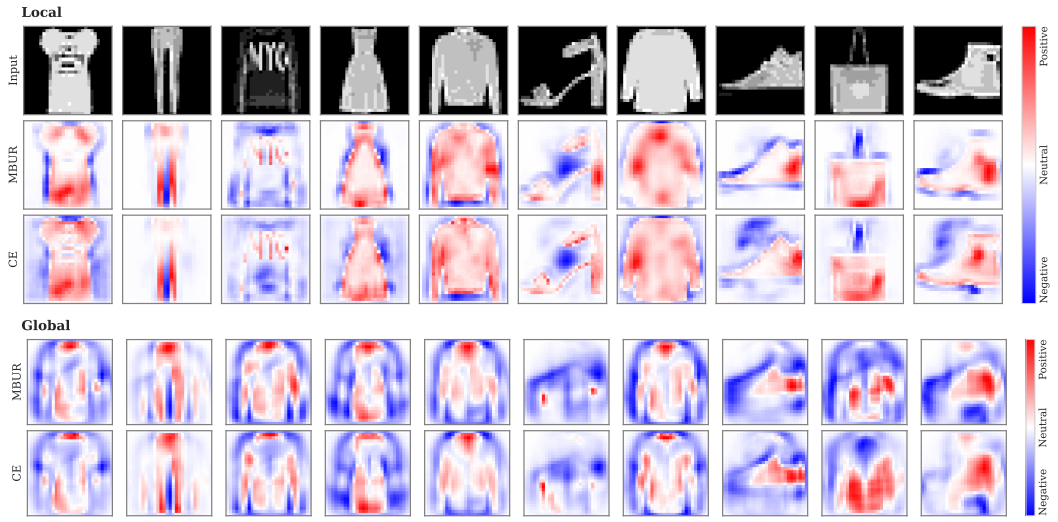


Figure 5: The local and global interpretation for the FashionMNIST dataset. Red indicates positive feature importance, while blue indicates negative importance. The color intensity indicates the relative magnitude of the importance.

to 100% while test accuracy plateaued between 81% and 86%. CE showed a substantially smaller train-test gap under the same sweep, with test accuracy consistently higher than MBUR's across every configuration tried (Figure 6). IMDB was the only dataset that received a more extensive hyperparameter search, due to the observed learning behavior of MBUR.

Table 4: Hyperparameters for the twelve IMDB sweep configurations. v12 is the configuration reported in the main text (Table 1).

					Test Accuracy	
Config	**Clauses**	s	T	η	MBUR	Loss-Guided CE
v1	2000	2.0	5000	0.1	0.813	0.837
v2	1000	2.0	5000	0.01	0.849	0.861
v3	500	2.0	5000	0.01	0.828	0.859
v4	500	1.1	5000	0.01	0.822	0.868
v5	10000	1.1	50000	0.01	0.857	0.875
v6	10000	1.1	10000	0.001	0.862	0.877
v7	20000	1.1	20000	0.001	0.862	0.877
v8	10000	1.1	10000	0.001	0.856	0.876
v9	10000	1.05	10000	0.001	0.854	0.869
v10	7000	1.1	10000	0.001	0.854	0.877
v11	10000	1.1	2000	0.001	0.857	0.876
v12	10000	1.1	10000	0.001	0.855	0.877

A.3 AG NEWS

Figure 7 shows the global interpretation for AG News using the same word-cloud procedure as IMDB, described in Section 5.1. MBUR finds a different set of top-magnitude words than CE and SCE, and for several classes these come from another class's vocabulary rather than the target class's own. For Sci/Tech, MBUR's most important word is "nasdaq" (1.00), a Business term, while "nasa" (-0.59), the canonical Sci/Tech term shows negative importance. CE and SCE both correct determines "nasa" (1.00) to be the most important and gives "nasdaq" near zero importance (0.06 and 0.02 respectively). For Sports, MBUR gives high importance to "analyst", "device", "phone",

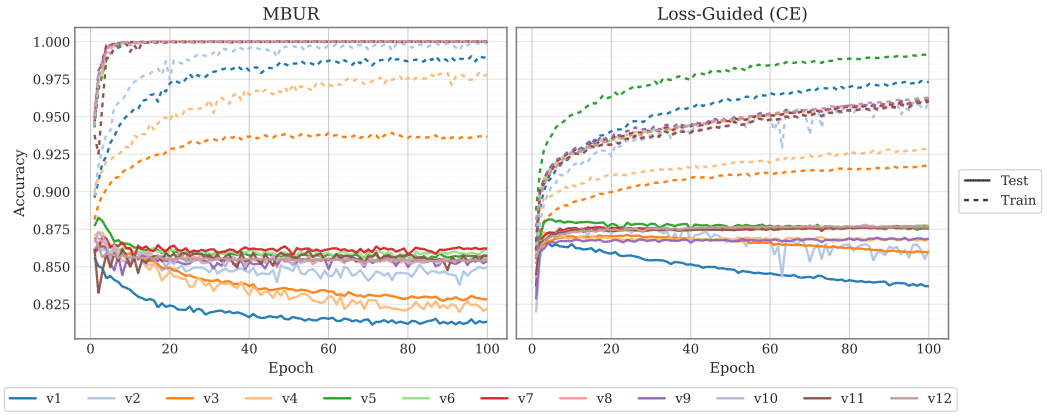


Figure 6: Training and test accuracy curves for all twelve IMDB sweep configurations (Table 4), using MBUR (left) and Loss-Guided CE (right). Solid curves denote test accuracy, dashed curves denote train accuracy.

and "crm" (0.79 to 0.90), all Business or Sci/Tech vocabulary, while "coach", the canonical Sports term is barely important (0.10). For the same class, CE and SCE correctly gives "coach" the maximum importance instead. For Business, MBUR gives high positive magnitude to "najaf", "falluja", "mosque", and "insurgency" (0.66 to 0.90), while "economy" and "treasury" have negative importance scores. CE and SCE again correct this, giving "economy" and "treasury" the highest importance instead. Despite the accuracy gap between MBUR and Loss-Guided being relatively small on AG News, Loss-Guided model's interpretation is more richer and accurate.

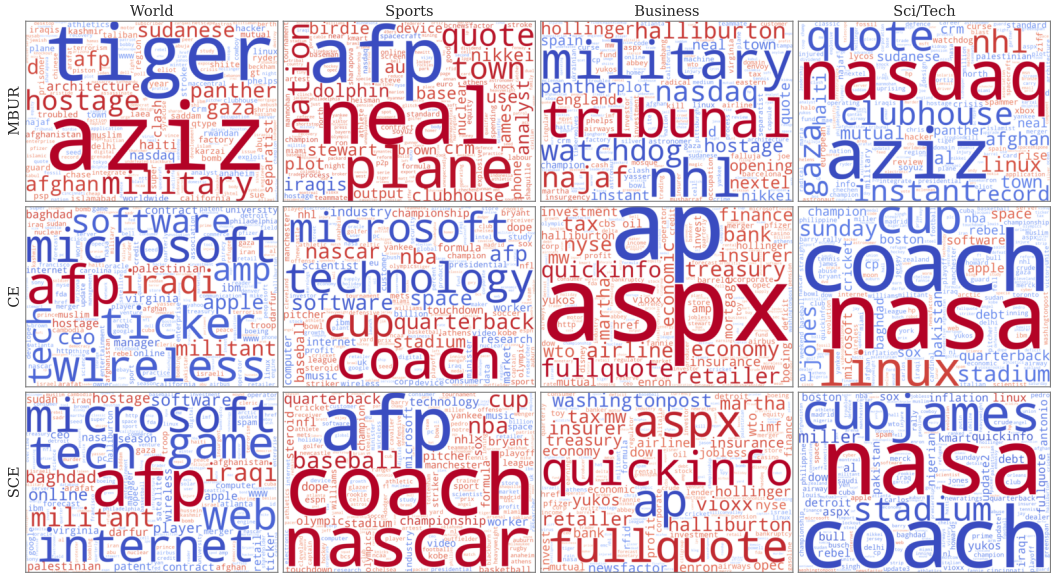


Figure 7: Global interpretation for AG News across MBUR, CE, and SCE. The red color indicates positive word importance, while the blue color indicates negative importance. The relative magnitude of the importance is shown using the size of the word and the intensity of the color.

A.4 GoEmotions

GoEmotions is a highly imbalanced multilabel dataset. We tried an inverse-frequency class-weighted cross entropy to compensate for this imbalance, but it gave no consistent improvement over plain cross entropy across the configurations tested.

Figure 8 shows the per-label frequency in the training set alongside the per-class test F1 for MBUR, CE, and ASL, both sorted by class rarity. Because the rarest classes have few training examples, none of the models perform reliably on them. Loss-Guided consistently improves over MBUR on the moderate to common classes.

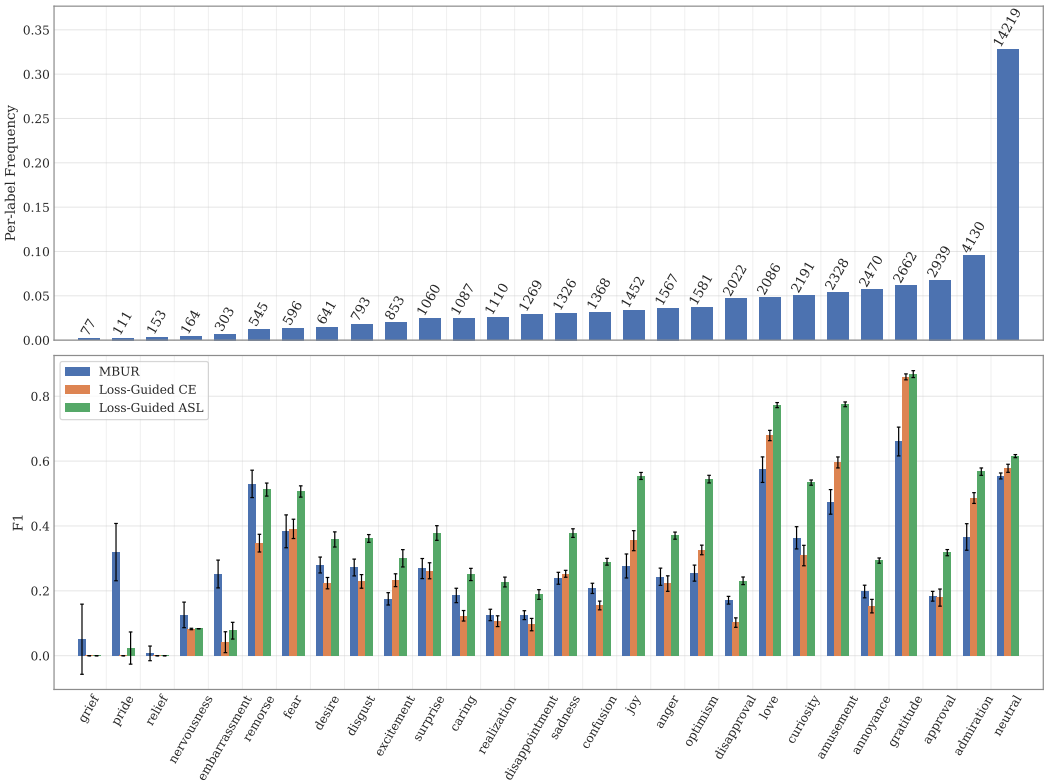


Figure 8: GoEmotions label distribution and performance sorted by class rarity from rarest to most common. The top panel shows training set frequency per label. The bottom panel displays test F1 scores (mean $\pm$ std over 10 seeds) for MBUR, CE, and ASL.

Figures 9 and 10 show the global interpretability for the five rarest and five most common classes. Unlike IMDB and AG News, the three conditions largely agree here, each surfacing similar, topically coherent vocabulary per class.

For the common classes, this agreement lines up with equally strong F1 for all three conditions. Gratitude is a typical case, "Thanks" and "Thank" dominate under all three conditions. The one exception among the ten classes shown is neutral, a catch-all class with no consistent topic, where all three conditions surface generic, uninterpretable words instead.

For the rarest classes, the vocabulary agreement holds even where the predictive performance does not. On grief, Loss-Guided's test F1 is 0.00 for both CE and ASL, while MBUR reaches 0.05 with high variance. On pride, CE is at 0.00 and ASL at 0.02, while MBUR reaches 0.32. Despite this, "die", "mourn", and "RIP" dominate grief's global interpretation under all three conditions, and "proud", "Proud", and "athletic" dominate pride's under all three as well. The model has learned the right vocabulary for these two classes even where Loss-Guided never predicts them. Relief, the rarest class with only 153 positive training examples, is weaker across the board. All three conditions score near zero F1 there, and their top words mix genuine relief vocabulary such as "relieved" and



Figure 9: Global interpretations for the five rarest GoEmotions classes.



Figure 10: Global interpretations for the five most common GoEmotions classes.

"glad" with non-informative tokens such as "goddd" and "red car", under MBUR as much as under CE and ASL.

A.5 CELEBA

CelebA's 40 attributes are imbalanced as well, though less severely than GoEmotions (Figure 11). As with GoEmotions, an inverse-frequency class-weighted cross entropy gave little to no improvement over plain cross entropy at matched hyperparameters. For ASL specific parameters, $\gamma_+ = \{0, 2\}$ and $\gamma_- = \{2, 4, 6\}$ were tested. $\gamma_+ = 2$ and $\gamma_- = 4$ gave the best results and are thus reported in the paper. Figure 11 shows the per-class test F1 for MBUR, CE, and ASL, sorted by class rarity. ASL's gain over MBUR here is significantly larger on the rarest attributes. The gains for more common labels are minimal.

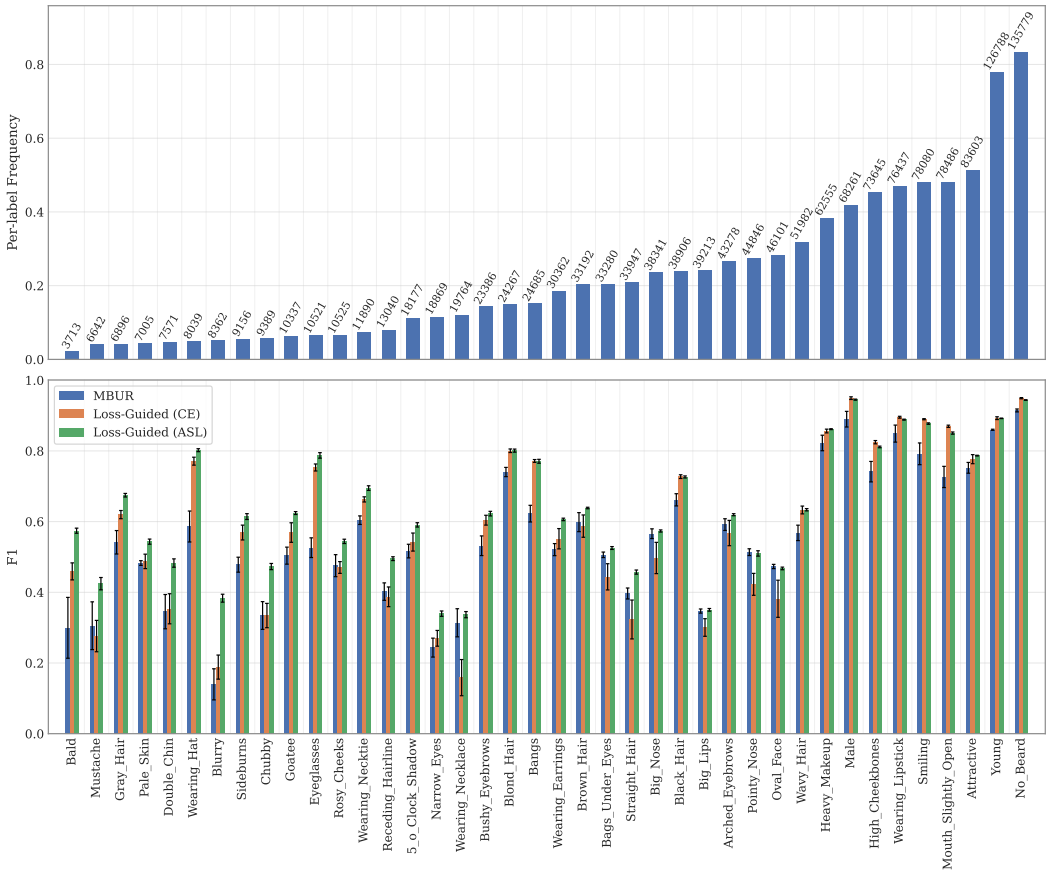


Figure 11: (Top) per-label frequency in the CelebA training set. (Bottom) per-class test F1 for MBUR, CE, and ASL, mean ± std over 10 seeds. Both sorted by class rarity (rarest first).