BEYOND FIXED EMOTION MAPPINGS: LEARNING PROBABILISTIC RELATIONS ACROSS EMOTION GRANULARITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion recognition is naturally hierarchical: fine-grained affective states such as *grief* or *annoyance* relate to broader categories such as *sadness* or *anger*. Existing emotion taxonomies specify such categorical relationships, but knowing the hierarchy does not uniquely determine how probability mass in a fine-grained predictive posterior should propagate into a coarse-grained distribution. We therefore formulate cross-granularity emotion prediction as learning a probabilistic relationship between emotion granularities while retaining known hierarchical structure.

Our hierarchical consistency framework jointly optimizes distinct fine- and coarse-grained predictors together with a learnable column-stochastic transformation M_θ. The transformation maps the complete fine-grained posterior to an induced coarse posterior, which is coupled with a directly predicted coarse posterior through distribution-level consistency. Structural regularization anchors M_θ to a soft empirical mapping M_{emp} derived exclusively from the multi-label training data.

Experiments on *GoEmotions* show that deterministic, empirical probabilistic, and learned probabilistic mappings yield nearly identical coarse-grained top-1 accuracy (78.26–78.36%) while inducing meaningfully different predictive distributions. Relative to the training-derived empirical mapping, a learned transformation achieves a 25.7% relative reduction in expected calibration error, while paired-bootstrap analysis supports reductions in negative log-likelihood for learned mappings. Importantly, the learned mappings preserve the dominant Ekman assignment of every fine-grained emotion. These results demonstrate that a known categorical hierarchy does not uniquely determine the probabilistic relationship across label granularities: mappings with nearly identical categorical behavior can exhibit systematically different predictive distributions.

1 INTRODUCTION

Emotion recognition in Natural Language Processing (NLP) is commonly formulated as classification over a fixed set of affective categories. Modern datasets increasingly provide fine-grained inventories; for example, *GoEmotions* Demszky et al. (2020) distinguishes 28 emotions, while psychological taxonomies organize related affective states into broader categories such as *anger*, *joy*, and *sadness* Ekman (1992); Plutchik (1980). This creates a general hierarchical prediction question: how should predictive distributions at one label granularity be related to those at another?

A conventional solution uses a fixed fine-to-coarse mapping. A hard mapping provides a clear categorical relationship but discards probability mass assigned to alternative fine-grained emotions. Probabilistic aggregation can instead propagate the complete fine-grained posterior, but a fixed transformation still prespecifies how that probability mass is distributed across granularities. Importantly, a taxonomy specifies categorical organization, whereas a predictive posterior represents input-dependent evidence over labels. The former therefore does not uniquely determine how the latter should be propagated. Mappings that preserve the same dominant fine-to-coarse associations may consequently produce similar top-1 predictions while inducing different coarse-grained probability distributions.

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.

We investigate whether these cross-granularity relations can instead be learned probabilistically while remaining anchored to known hierarchical structure. We propose a **Hierarchical Consistency Framework** with distinct fine- and coarse-grained predictors and a learnable column-stochastic transformation M_θ. The transformation maps the complete fine-grained posterior to an induced coarse posterior, which is coupled with a directly predicted coarse posterior through distribution-level consistency. A soft empirical mapping M_{emp}, constructed exclusively from multi-label training data, provides structural regularization without fixing the transformation.

On *GoEmotions*, deterministic, empirical probabilistic, and learned probabilistic mappings yield similar coarse-grained top-1 accuracies (78.26–78.36%), and the learned mappings preserve the dominant predefined coarse assignment for every fine-grained emotion. Their predictive distributions nevertheless differ. Relative to M_{emp}, a learned transformation achieves a 25.7% relative reduction in expected calibration error, while paired-bootstrap analysis supports reductions in negative log-likelihood for learned mappings. Thus, learning primarily affects the distribution of probability mass rather than the dominant coarse prediction.

We investigate three research questions:

- **RQ1: Categorical structure vs. probabilistic propagation.** How does probabilistic propagation of the complete fine posterior differ from deterministic hierarchical projection?
- **RQ2: Empirical vs. learned relations.** How does learning the cross-granularity transformation affect coarse posterior quality relative to a training-derived empirical mapping when top-1 predictions are similar?
- **RQ3: Consistency vs. structural regularization.** How do these constraints affect predictive agreement, mapping structure, and performance?

Our contributions are threefold:

1. We formulate cross-granularity emotion prediction as learning a **probabilistic relationship between emotion granularities**, separating categorical organization from posterior propagation.
2. We introduce a **Hierarchical Consistency Framework** combining distinct fine- and coarse-grained predictors, a learnable column-stochastic mapping M_θ, distribution-level consistency, and training-derived structural regularization.
3. We compare deterministic, empirical, and learned mappings using decision-level, structural, probability-quality, and paired-bootstrap analyses, showing that similar top-1 behavior can coexist with systematic differences in predictive distribution quality.

2 Related Work

Fine-grained and hierarchical emotion recognition. Emotion recognition has traditionally been influenced by psychological theories that organize affect into broad categories Ekman (1992); Plutchik (1980); Shaver et al. (1987), while recent datasets provide substantially finer emotion inventories. *GoEmotions* Demszky et al. (2020), for example, contains 27 emotion categories plus Neutral and also considers mappings to broader sentiment and Ekman groupings. Transformer models such as BERT Devlin & et al. (2018) provide effective representations for learning these fine-grained distinctions. However, standard emotion classification generally operates within a single label space and does not explicitly model how a predictive posterior at one granularity should be related to a posterior at another.

Hierarchical classification exploits relationships among labels rather than treating them as structurally independent Silla & Freitas (2011). Multi-task learning similarly supports joint prediction objectives, although interactions among tasks can affect optimization Standley et al. (2020). Joint prediction at multiple levels, however, does not by itself specify a probabilistic transformation between their output spaces. Our setting instead makes the fine-to-coarse predictive relationship an explicit component of the learning problem.

Cross-granularity mappings. A predefined hierarchy specifies categorical relationships between fine- and coarse-grained labels, but it does not uniquely determine how probability mass in a predic-

tive posterior should be propagated across those labels. A deterministic mapping uses the dominant fine-grained prediction, whereas a probabilistic mapping can aggregate the complete fine-grained posterior. Learned hierarchical-transition approaches explicitly model cross-level label relationships. Wang et al. Wang et al. (2026) learn label-hierarchy transition matrices through a transition network, while fine-to-coarse mapping matrices have been used in fine-grained named entity recognition to exploit coarse-grained supervision Lee et al. (2023).

Rather than treating the cross-level transformation primarily as a mechanism for improving hierarchical classification, our work studies the probabilistic relationship itself. We distinguish a predefined deterministic mapping M_{hard}, a training-derived empirical mapping M_{emp}, and a jointly learned transformation M_θ. The learned transformation is structurally anchored to M_{emp} and coupled to fine- and coarse-grained predictions through distribution-level consistency. This formulation enables a controlled comparison of whether mappings that preserve the same dominant hierarchical relationships can nevertheless induce different predictive distributions.

Probability quality and calibration. Accuracy and F1 characterize hard classification decisions but not the quality of the complete predictive distribution. Proper scoring rules such as negative log-likelihood and Brier score provide distribution-sensitive evaluation Brier (1950); Gneiting & Raftery (2007), while calibration measures whether predictive confidence corresponds to empirical correctness. Modern neural networks can remain poorly calibrated despite strong classification performance Guo et al. (2017), and recent work has investigated calibration specifically for emotion detection Petre-Vlad et al. (2025). We therefore evaluate cross-granularity mappings using both hard-decision metrics and distribution-sensitive measures, including NLL, Brier score, and ECE, to determine whether similar categorical predictions conceal differences in probability quality.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND ARCHITECTURE

We consider emotion prediction at two related granularities. Let $\mathcal{F} = \{1, \ldots, K_f\}$ and $\mathcal{C} = \{1, \ldots, K_c\}$ denote the fine- and coarse-grained emotion spaces, respectively, with $K_c < K_f$. Given an input x, two independently parameterized encoders produce

$$p^f(x) = \text{softmax}(W_f f_{\theta_f}(x)), \qquad p^c(x) = \text{softmax}(W_c f_{\theta_c}(x)), \tag{1}$$

where $p^f(x) \in \Delta^{K_f - 1}$ and $p^c(x) \in \Delta^{K_c - 1}$. The fine- and coarse-grained branches do not share encoder parameters.

As illustrated in Figure 1, we additionally learn a cross-granularity transformation $M_\theta \in [0, 1]^{K_c \times K_f}$ that maps the complete fine-grained posterior to an induced coarse posterior,

$$\hat{p}^c(x) = M_\theta p^f(x). \tag{2}$$

Each column of M_θ sums to one and represents a distribution over coarse emotions associated with a fine-grained emotion. The framework therefore provides two independently parameterized paths to the coarse space: direct prediction $p^c(x)$ and hierarchical prediction $\hat{p}^c(x)$.

3.2 LEARNING CROSS-GRANULARITY RELATIONS

We parameterize each column of M_θ as

$$(M_\theta)_{j,i} \equiv P_\theta(C = j \mid F = i), \qquad \sum_{j=1}^{K_c} (M_\theta)_{j,i} = 1. \tag{3}$$

A learnable matrix $A \in \mathbb{R}^{K_c \times K_f}$ is constrained and normalized during each forward pass:

$$\widetilde{A}_{j,i} = \text{clip}(A_{j,i}, \epsilon, 1), \qquad (M_\theta)_{j,i} = \frac{\widetilde{A}_{j,i}}{\sum_{k=1}^{K_c} \widetilde{A}_{k,i}}. \tag{4}$$

Consequently,

$$\hat{p}^c_j(x) = \sum_{i=1}^{K_f} (M_\theta)_{j,i} p^f_i(x), \tag{5}$$

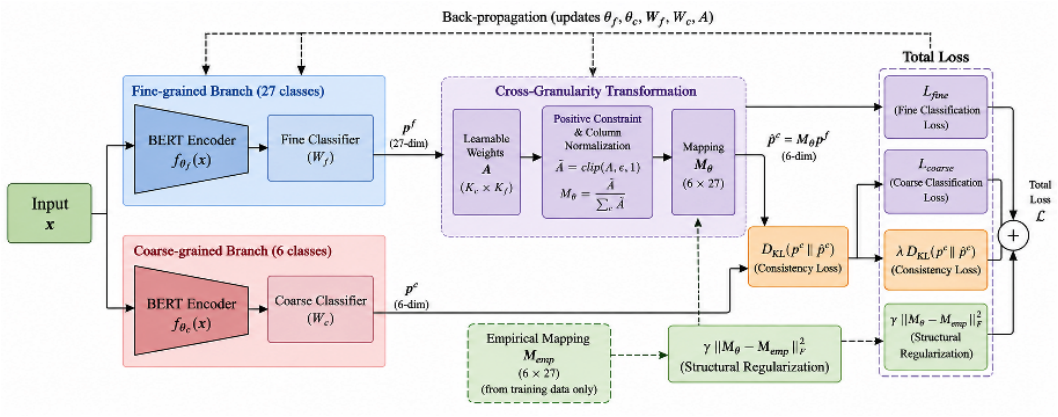


Figure 1: Hierarchical consistency framework. Independently parameterized encoders produce p^f and p^c. The column-stochastic transformation M_θ maps p^f to $\hat{p}^c = M_\theta p^f$. Distribution-level consistency couples p^c and $\hat{p}^c$, while structural regularization anchors M_θ to the training-derived mapping M_{emp}.

so coarse prediction aggregates probability mass from the complete fine-grained posterior rather than only its most probable class.

3.3 Consistency and Structural Regularization

Given fine- and coarse-grained labels y^f and y^c, the supervised losses are

$$\mathcal{L}_f = \text{CE}(y^f, p^f(x)), \qquad \mathcal{L}_c = \text{CE}(y^c, p^c(x)). \tag{6}$$

We couple the direct and hierarchy-induced coarse posteriors through

$$\mathcal{L}_{\text{cons}} = D_{\text{KL}}(p^c(x) \| \hat{p}^c(x)) \,. \tag{7}$$

Because $\hat{p}^c(x) = M_\theta p^f(x)$, this objective couples the learned mapping and fine-grained predictor with the independently parameterized direct coarse predictor.

We distinguish three cross-granularity transformations. M_{hard} is the deterministic one-hot mapping defined by the predefined fine-to-coarse Ekman assignments; M_{emp} is a soft empirical mapping estimated exclusively from the training partition; and M_θ is learned jointly with the prediction models.

To construct M_{emp}, we preserve the original multi-label training annotations, map their annotated fine emotions to predefined Ekman categories, accumulate fine–coarse co-occurrences, and normalize each fine-emotion column:

$$(M_{\text{emp}})_{j,i} = \frac{N_{\text{train}}(F = i, C = j)}{\sum_{k=1}^{K_c} N_{\text{train}}(F = i, C = k)}, \qquad \sum_{j=1}^{K_c} (M_{\text{emp}})_{j,i} = 1. \tag{8}$$

Thus, M_{emp} can retain soft associations arising from multi-label co-occurrence even though M_{hard} assigns each fine emotion to a single predefined category. No validation or test labels are used in its construction.

Rather than fixing M_θ to this empirical relationship, we use M_{emp} as a structural reference:

$$\mathcal{L}_{\text{struct}} = \|M_\theta - M_{\text{emp}}\|_F^2. \tag{9}$$

3.4 Joint Learning Objective

The complete objective is

$$\mathcal{L} = \mathcal{L}_f + \mathcal{L}_c + \lambda \mathcal{L}_{\text{cons}} + \gamma \mathcal{L}_{\text{struct}}, \tag{10}$$

Table 1: Fine- and coarse-grained prediction performance on the held-out GoEmotions test set.

Prediction path	Accuracy	Macro-F1
Fine classifier (27 classes)	0.5478	0.4733
Direct coarse classifier (6 classes)	0.7837	0.6643
Hard hierarchical projection	0.7826	0.6582
Learned hierarchical mapping ($\gamma = 0.5$)	0.7834	0.6551

where λ controls prediction-level consistency and γ controls structural regularization toward M_{emp}. All parameters $\Theta = \{\theta_f, \theta_c, W_f, W_c, A\}$ are optimized jointly by back-propagation. The complete optimization procedure is summarized in Algorithm 1 in Appendix A.

The formulation separates categorical structure from probabilistic cross-granularity relations: M_{hard} specifies the predefined dominant association, M_{emp} captures soft relations observed in the multi-label training data, and M_θ adapts these relations through joint learning while remaining structurally constrained.

4 Experiments

4.1 Experimental Setup

We evaluate the proposed framework on GoEmotions using the 27 non-neutral fine-grained emotion categories and six coarse-grained Ekman categories: anger, disgust, fear, joy, sadness, and surprise. Original instances are partitioned before multi-label examples are expanded into single-label examples, preventing labels associated with the same utterance from entering different data splits.

The fine-grained and direct coarse-grained predictors use separate BERT-base-uncased encoders, producing a 27-dimensional posterior $p^f(x)$ and a six-dimensional posterior $p^c_{\text{direct}}(x)$, respectively. The learned transformation $M_\theta \in [0, 1]^{6\times 27}$ maps the complete fine-grained posterior into the coarse space. The empirical mapping M_{emp} is estimated exclusively from the training partition; no validation or test labels are used in its construction. Unless otherwise specified, $\lambda = 0.5$ and $\gamma = 0.5$.

Models are trained for up to 10 epochs with batch size 16, maximum sequence length 128, BERT learning rate 2×10^{-5}, mapping learning rate 10^{-3}, and random seed 42. Checkpoints are selected by validation mapped Macro-F1, and the held-out test set is used only for final evaluation.

We report accuracy and Macro-F1 for top-1 classification and NLL, Brier score, and ECE (15 bins) for probability quality. Statistical reliability is assessed using paired bootstrap resampling with 10,000 resamples.

4.2 Overall Hierarchical Prediction Performance

We first evaluate prediction at the two emotion granularities. Table 1 reports performance on the held-out test set of 9,192 instances.

The direct six-class predictor achieves higher classification performance than the 27-class predictor, reflecting the lower ambiguity of the coarser prediction problem. Importantly, hierarchical prediction from the fine-grained posterior recovers nearly the same coarse accuracy: the learned hierarchical path achieves 0.7834, compared with 0.7837 for the independently trained direct coarse classifier. We next isolate the effect of the cross-granularity transformation by applying alternative mappings to the same fine-grained posterior.

4.3 Hard, Empirical, and Learned Cross-Granularity Mappings

We compare three forms of cross-granularity structure: the deterministic taxonomy M_{hard}, the training-derived probabilistic mapping M_{emp}, and the learned mapping M_θ. For the learned mapping, we vary γ while fixing $\lambda = 0.5$ to examine the effect of structural regularization. All mappings are applied to the same frozen fine-grained posterior $p^f(x)$. Table 2 reports both predictive performance and structural properties of the mappings.

Table 2: Comparison of hard, empirical, and learned fine-to-coarse mappings on the held-out test set. Learned mappings use $\lambda = 0.5$ while varying γ. $\bar{H}(M)$ is the mean column entropy, and $d_1(M, M')$ is the mean column-wise L_1 distance from the indicated reference mapping.

Mapping M	Acc.	Macro-F1	$\bar{H}(M)$	$d_1(M, M_{\text{hard}})$	$d_1(M, M_{\text{emp}})$
M_{hard}	0.7826	**0.6582**	0.0000	0.0000	0.2786
M_{emp}	0.7833	0.6554	0.5564	0.2786	0.0000
M_θ ($\gamma = 0$)	**0.7836**	0.6546	0.4568	0.2331	0.0592
M_θ ($\gamma = 0.5$)	0.7834	0.6551	0.5250	0.2649	0.0257
M_θ ($\gamma = 1.0$)	0.7834	0.6561	0.5402	0.2725	0.0168

For a mapping M, we characterize its softness using the mean column entropy

$$\bar{H}(M) = -\frac{1}{K_f} \sum_{i=1}^{K_f} \sum_{j=1}^{K_c} M_{j,i} \log M_{j,i}, \tag{11}$$

and its mean column-wise L_1 distance from a reference mapping M',

$$d_1(M, M') = \frac{1}{K_f} \sum_{i=1}^{K_f} \left\| M_{:,i} - M'_{:,i} \right\|_1 . \tag{12}$$

We use M_{hard} and M_{emp} as the two reference mappings.

As shown in Table 2, top-1 performance is highly stable across the five mappings. Accuracy ranges from 0.7826 to 0.7836, while Macro-F1 ranges from 0.6546 to 0.6582. M_{hard} obtains the highest Macro-F1, whereas the learned mapping with $\gamma = 0$ obtains the highest accuracy. The small and non-uniform differences indicate that replacing the hard taxonomy with a probabilistic mapping does not substantially change coarse-grained argmax classification.

The mappings show larger differences in their structural properties. M_{hard} has zero mean column entropy, whereas M_{emp} has $\bar{H}(M_{\text{emp}}) = 0.5564$. For the learned mappings, increasing γ from 0 to 1.0 reduces $d_1(M_\theta, M_{\text{emp}})$ from 0.0592 to 0.0168 and increases $\bar{H}(M_\theta)$ from 0.4568 to 0.5402. Stronger structural regularization therefore moves M_θ toward the soft empirical structure, as intended.

Importantly, the maximum entry in every learned mapping column corresponds to the same coarse category as in M_{hard}. The learned mappings therefore redistribute probability mass across coarse categories while preserving the dominant predefined taxonomic relation. This distinction motivates the probability-quality analysis in the following subsection. Detailed instance-level and class-specific disagreement analyses are provided in Appendix B.1.

4.4 Probability Quality of Cross-Granularity Predictions

The preceding results show that the mappings produce nearly identical top-1 accuracies despite differences in probabilistic structure. We therefore evaluate the complete coarse-grained predictive distributions using NLL, Brier score, and ECE with 15 confidence bins. Lower values indicate better performance; mean confidence is reported descriptively.

Table 3 reveals differences largely hidden by top-1 accuracy. Hard projection has the highest mean confidence (0.8637) but also the largest NLL, Brier score, and ECE. Replacing hard projection with probabilistic aggregation reduces Brier score from 0.3211 to approximately 0.314 and ECE from 0.0811 to 0.0342 or lower, while leaving accuracy nearly unchanged.

Learning the cross-granularity transformation produces additional differences relative to M_{emp}. All three learned mappings obtain lower NLL, while the $\gamma = 0.5$ configuration further reduces ECE from 0.0342 to 0.0254, a 25.7% relative reduction, and slightly reduces Brier score from 0.3139 to 0.3136. No single γ dominates all probability metrics. These results indicate that the principal effect of probabilistic cross-granularity learning is the redistribution of probability mass rather than alteration of the dominant coarse label.

Table 3: Probability quality and calibration of coarse-grained predictions on the 9,192 test instances. Lower NLL, Brier score, and ECE are better. ECE is computed using 15 confidence bins; Mean Conf. denotes the average maximum predicted class probability.

Mapping	Acc.	NLL ↓	Brier ↓	ECE ↓	Mean Conf.
Hard M_{hard}	0.7826	0.6477	0.3211	0.0811	0.8637
Empirical M_{emp}	0.7833	0.6435	0.3139	0.0342	0.7723
Learned ($\gamma = 0$)	**0.7836**	**0.6340**	0.3140	0.0279	0.8040
Learned ($\gamma = 0.5$)	0.7834	0.6381	**0.3136**	**0.0254**	0.7823
Learned ($\gamma = 1.0$)	0.7834	0.6411	**0.3136**	0.0301	0.7777

Table 4: Paired-bootstrap comparisons between learned and empirical mappings over 10,000 resamples. For each pair, $\Delta = \mathcal{L}_{\text{learned}} - \mathcal{L}_{\text{emp}}$; negative values favor the learned mapping. The 95% CIs are for the pairwise differences.

Metric	Learned mapping	Δ	95% CI
NLL	$\gamma = 0$	-0.00946	$[-0.01195, -0.00694]$
NLL	$\gamma = 0.5$	-0.00545	$[-0.00635, -0.00450]$
NLL	$\gamma = 1.0$	-0.00244	$[-0.00329, -0.00150]$
Brier	$\gamma = 0$	$+0.00006$	$[-0.00049, +0.00061]$
Brier	$\gamma = 0.5$	-0.00038	$[-0.00058, -0.00017]$
Brier	$\gamma = 1.0$	-0.00035	$[-0.00046, -0.00024]$

4.5 Statistical Reliability of Probability-Quality Differences

We assess the reliability of the NLL and Brier differences using paired bootstrap resampling of the 9,192 test instances with 10,000 resamples. For each learned mapping, we compute $\Delta = \mathcal{L}_{\text{learned}} - \mathcal{L}_{\text{emp}}$, where $\mathcal{L}$ denotes NLL or Brier score; negative values therefore favor the learned mapping. Table 4 reports the observed differences and their paired-bootstrap 95% confidence intervals.

As shown in Table 4, all three learned mappings reduce NLL relative to M_{emp}, with 95% CIs excluding zero. For Brier score, the intervals for $\gamma = 0.5$ and $\gamma = 1.0$ lie below zero, whereas the $\gamma = 0$ interval contains zero. Thus, learning provides consistent NLL improvements over the fixed empirical transformation, with smaller Brier improvements for the structurally regularized configurations.

Additional paired-bootstrap comparisons against M_{hard} show that all probabilistic mappings yield lower Brier scores with 95% CIs excluding zero. For NLL, the intervals exclude zero for the learned $\gamma = 0$ and $\gamma = 0.5$ mappings, whereas those for M_{emp} and learned $\gamma = 1.0$ include zero. These results separate the effect of probabilistic aggregation from the additional effect of learning M_θ.

4.6 Sensitivity to Consistency and Structural Regularization

We finally examine the roles of λ and γ, varying one weight at a time while holding the other fixed. All results in this subsection are reported on the validation set using checkpoints selected by validation mapped Macro-F1; the test set is not used for hyperparameter selection.

Prediction-consistency weight. Table 5 varies $\lambda \in \{0, 0.1, 0.5, 1.0\}$ with $\gamma = 0.5$.

As shown in Table 5, increasing λ reduces validation KL from 0.1276 to 0.0553, a 56.6% reduction, while Macro-F1 and accuracy remain comparatively stable. This supports the intended role of λ as a prediction-level consistency constraint rather than primarily as a mechanism for improving classification accuracy.

Structural regularization weight. Table 6 varies $\gamma \in \{0, 0.5, 1.0\}$ with $\lambda = 0.5$.

Table 5: Sensitivity to the prediction-consistency weight λ with $\gamma = 0.5$. KL denotes validation $D_{\mathrm{KL}}(p^c \| \hat{p}^c)$.

λ	Mapped Macro-F1	Mapped Acc.	Val. KL $\downarrow$
0.0	0.6568	0.7812	0.1276
0.1	0.6585	0.7793	0.1215
0.5	0.6579	0.7830	0.0791
1.0	0.6616	0.7826	0.0553

Table 6: Sensitivity to the structural regularization weight γ with $\lambda = 0.5$. Structural distances are measured relative to the training-derived empirical mapping M_{emp}.

γ	Macro-F1	$\|M_\theta - M_{\mathrm{emp}}\|_F^2$	$d_1(M_\theta, M_{\mathrm{emp}})$	$\bar{H}(M_\theta)$
0.0	0.6575	0.03835	0.05925	0.4568
0.5	0.6579	0.00711	0.02569	0.5250
1.0	0.6637	0.00393	0.01681	0.5402
M_{emp}	–	0.00000	0.00000	0.5564

Table 6 shows the expected structural effect: increasing γ reduces $\|M_\theta - M_{\mathrm{emp}}\|_F^2$ from 0.03835 to 0.00393 and $d_1(M_\theta, M_{\mathrm{emp}})$ from 0.05925 to 0.01681, while $\bar{H}(M_\theta)$ approaches that of M_{emp}. Mapped Macro-F1 varies only modestly across the three settings, and the dominant predefined coarse assignment is preserved for every fine emotion at all three values of γ.

Together, the sensitivity analyses distinguish the two regularizers: λ controls prediction-level agreement between the direct and hierarchical coarse paths, whereas γ controls structural adherence of M_θ to the empirical cross-granularity relationship.

5 Discussion

Our results highlight a distinction between categorical hierarchical structure and probabilistic relationships across emotion granularities. A predefined hierarchy specifies the dominant coarse category associated with each fine-grained emotion, but does not uniquely determine how the complete fine-grained posterior should be propagated to the coarse space. This distinction explains why the deterministic, empirical, and learned mappings yield nearly identical coarse-grained top-1 performance while producing different predictive distributions.

What is learned beyond a fixed hierarchy? The three mappings encode progressively different assumptions. M_{hard} specifies categorical fine-to-coarse membership, whereas M_{emp} captures soft associations observed in the multi-label training data. M_θ further allows these probabilistic relationships to adapt jointly with the predictive models, with γ controlling the degree to which the learned transformation is anchored to the empirical structure. Learning M_θ is therefore not intended primarily to discover a different dominant hierarchy, but to determine how probability mass is distributed across coarse categories. Consistent with this interpretation, the learned mappings preserve the dominant predefined coarse assignment for every fine emotion while modifying their non-dominant probability mass.

Decision-level similarity can conceal distributional differences. Accuracy and Macro-F1 alone make the mapping strategies appear largely interchangeable. Distribution-sensitive metrics reveal differences that hard decisions conceal. Relative to M_{emp}, the structurally regularized learned mapping reduces ECE from 0.0342 to 0.0254, a 25.7% relative reduction, while all three learned mappings reduce NLL with paired-bootstrap 95% confidence intervals excluding zero. Probabilistic aggregation also improves Brier score relative to deterministic projection. These results should not be interpreted as substantial improvements in coarse classification accuracy; rather, they show that mappings with nearly identical categorical decisions can differ in how probability is allocated

among competing coarse emotions. Such differences matter when downstream models or decision procedures consume predictive probabilities rather than only argmax labels.

Complementary constraints and broader implications. The sensitivity results clarify the distinct roles of the two regularizers. Increasing λ strengthens prediction-level agreement between the direct coarse posterior p^c and the hierarchy-induced posterior $\hat{p}^c$, whereas increasing γ anchors M_θ more strongly to M_{emp}. Thus, λ regulates agreement between predictive paths, while γ regulates the structure of the learned transformation.

More generally, knowing that a fine category belongs to a particular parent does not fully specify how uncertainty over fine categories should propagate to a distribution over parent categories. Hierarchical prediction should therefore be evaluated at both the decision and distribution levels. A learned transformation need not replace a known hierarchy to be useful: it can preserve its dominant categorical organization while adapting the softer probabilistic relationships through which predictive evidence is propagated across granularities.

6 Limitations and Future Work

Our study has several limitations. First, evaluation is limited to *GoEmotions* and an Ekman-based coarse emotion space, and the empirical mapping M_{emp} reflects co-occurrence patterns in the corresponding training annotations. Cross-granularity relationships may therefore depend on both the dataset and taxonomy. Future work should evaluate additional emotion datasets, alternative taxonomies, and hierarchical classification domains, as well as structural guidance derived from external resources or independently specified relations.

Second, M_θ is a global transformation shared across instances. Although input-specific information is represented by $p^f(x)$, the mapping itself does not vary with linguistic context. Future work could investigate input-conditioned transformations $M_\theta(x)$ and examine whether their additional flexibility improves probability quality while retaining interpretable hierarchical structure.

Third, probability quality is multidimensional. We evaluate NLL, Brier score, and ECE, but ECE depends on the chosen binning procedure and no single metric fully characterizes predictive distributions. Future work could consider additional calibration measures, alternative binning strategies, and post-hoc calibration methods.

Theoretical direction. Our empirical findings reveal that different cross-granularity mappings can preserve the same dominant categorical relationships while inducing different predictive distributions. This observation motivates further theoretical study of cross-granularity probability propagation. In particular, future work could characterize conditions under which different mappings preserve the same argmax decisions and derive bounds relating changes in the mapping to changes in the induced predictive distribution.

7 Conclusion

We investigated cross-granularity emotion prediction when the categorical hierarchy is known but the probabilistic relationships between granularities are not fixed. Our hierarchical consistency framework jointly learns distinct fine- and coarse-grained predictors with a column-stochastic transformation M_θ, using prediction-level consistency and structural regularization toward a training-derived empirical mapping M_{emp}.

On *GoEmotions*, deterministic, empirical, and learned mappings produce nearly identical top-1 behavior while inducing different predictive distributions. The learned mappings preserve the dominant predefined coarse assignment for every fine-grained emotion, while a structurally regularized configuration reduces ECE from 0.0342 to 0.0254 relative to M_{emp} (25.7% relative), and all three learned mappings reduce NLL with paired-bootstrap 95% confidence intervals excluding zero. These results show that a known categorical hierarchy does not uniquely determine how predictive probability should propagate across granularities: learning can preserve the dominant hierarchy while adapting its softer probabilistic relationships.

AI USE STATEMENT

ChatGPT (OpenAI) and Gemini (Google) were used during the development of this work to assist with research methodology and experimental design, interpretation and organization of experimental results, refinement of the presentation of mathematical formulations, and drafting and editing portions of the manuscript. These generative AI tools were also used to provide feedback on scientific figures and the organization and clarity of the paper.

All model designs, experimental configurations, implementations, executions of experiments, data analyses, and final scientific claims were reviewed and verified by the authors. AI-generated suggestions were not treated as experimental evidence or as substitutes for validation of the reported results. The authors take full responsibility for the methodology, results, claims, citations, and final content of the paper.

ETHICS STATEMENT

This work uses the publicly available *GoEmotions* dataset and does not involve the collection of new human-subject data or interaction with human participants. Our experiments concern computational modeling of relationships between emotion label spaces. We recognize that emotion recognition models may reflect biases present in their training data and that probabilistic predictions of emotion should not be interpreted as definitive assessments of an individual's internal emotional state. Applications of such models in sensitive or high-stakes settings should therefore consider dataset limitations, potential demographic and contextual biases, and the uncertainty associated with model predictions.

REPRODUCIBILITY STATEMENT

We provide the methodological details required to reproduce the proposed framework in Section 3, including the fine- and coarse-grained prediction branches, construction of the cross-granularity transformation, consistency objective, empirical mapping, structural regularization, and joint learning objective. Section 4 describes the dataset processing, experimental protocol, model configuration, optimization settings, evaluation metrics, and model-selection procedure. The experiments include controlled comparisons among deterministic, empirical probabilistic, and learned probabilistic mappings, together with sensitivity analyses and paired-bootstrap evaluation. Additional implementation details and supporting experimental results are provided in the supplementary material. Code necessary to reproduce the experiments will be included with the anonymous supplementary submission.

REFERENCES

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

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 (ACL)*, 2020.

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

Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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

Algorithm 1 Joint Learning of Cross-Granularity Emotion Relations

Require: Training instances (x, y^f, y^c), training-derived empirical mapping M_{emp}, weights λ and γ

Ensure: Optimized parameters $\Theta = \{\theta_f, \theta_c, W_f, W_c, A\}$

Initialize fine and coarse encoders, prediction heads, and learnable mapping weights A with small positive values
while not converged **do**
  Sample a mini-batch (x, y^f, y^c)
  $h_f \leftarrow f_{\theta_f}(x)$
  $h_c \leftarrow f_{\theta_c}(x)$
  $p^f \leftarrow \text{softmax}(W_f h_f)$
  $p^c \leftarrow \text{softmax}(W_c h_c)$
  $\widetilde{A} \leftarrow \text{clip}(A, \epsilon, 1)$
  $M_\theta \leftarrow \text{ColumnNormalize}(\widetilde{A})$
  $\hat{p}^c \leftarrow M_\theta p^f$
  $\mathcal{L}_f \leftarrow \text{CE}(y^f, p^f)$
  $\mathcal{L}_c \leftarrow \text{CE}(y^c, p^c)$
  $\mathcal{L}_{\text{cons}} \leftarrow D_{\text{KL}}(p^c \| \hat{p}^c)$
  $\mathcal{L}_{\text{struct}} \leftarrow \|M_\theta - M_{\text{emp}}\|_F^2$
  $\mathcal{L} \leftarrow \mathcal{L}_f + \mathcal{L}_c + \lambda \mathcal{L}_{\text{cons}} + \gamma \mathcal{L}_{\text{struct}}$
  Update $\Theta = \{\theta_f, \theta_c, W_f, W_c, A\}$ by back-propagating $\mathcal{L}$
end while
return Θ

Su Lee, Seokjin Oh, and Woohwan Jung. Enhancing low-resource fine-grained named entity recognition by leveraging coarse-grained datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3269–3279, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.197.

Mihaela Petre-Vlad, Cornelia Caragea, and Florentina Hristea. Model calibration for emotion detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20442–20457, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1114.

Robert Plutchik. *A General Psychoevolutionary Theory of Emotion*. Academic Press, New York, 1980.

Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'Connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086, 1987.

Ricardo N Silla and Alex A Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22:31–72, 2011.

Trevor Standley et al. Which tasks should be learned together in multi-task learning? In *ICML*, 2020.

Renzhen Wang, De Cai, Kaiwen Xiao, Xixi Jia, Xiao Han, and Deyu Meng. Label hierarchy transition: Delving into class hierarchies to enhance deep classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):10112–10128, 2026. doi: 10.1109/TPAMI.2026.3683788.

A ALGORITHM

Algorithm 1 summarizes the joint optimization procedure.

Table 7: Pairwise disagreement analysis among the hard, empirical, and learned ($\gamma = 1.0$) mappings on the 9,192 test instances. "A only" and "B only" denote disagreement cases for which only the corresponding mapping predicts the correct coarse label; "Net (B)" is the difference between the numbers of B-only and A-only cases.

Comparison	Disagree	A only	B only	Net (B)
M_{hard} vs. M_{emp}	136	50	56	+6
M_{hard} vs. M_θ ($\gamma = 1.0$)	134	49	56	+7
M_{emp} vs. M_θ ($\gamma = 1.0$)	2	0	1	+1

B Additional Experimental Analysis

B.1 Analysis of Mapping Disagreements

The preceding results show only small differences in aggregate top-1 performance among the hard, empirical, and learned mappings. We therefore examine the instances for which these mappings produce different coarse-grained predictions. This analysis characterizes how probabilistic aggregation changes the decisions induced by the hard taxonomy and whether the empirical and learned transformations produce similar changes. For the learned transformation, we use the $\gamma = 1.0$ configuration, which is closest to M_{emp} among the learned mappings considered above.

As shown in Table 7, M_{hard} and M_{emp} disagree on 136 instances (1.48%), while M_{hard} and M_θ disagree on 134 (1.46%). In contrast, M_{emp} and M_θ disagree on only two instances (0.02%). Among the M_{hard}–M_θ disagreements, the learned mapping is uniquely correct on 56 instances, whereas the hard mapping is uniquely correct on 49, yielding a net difference of seven predictions. The corresponding comparison between M_{hard} and M_{emp} yields a net difference of six. Thus, probabilistic aggregation modifies only a small fraction of the decisions produced by the deterministic mapping, while the empirical and learned probabilistic mappings yield nearly identical argmax predictions.

The effects of these changes are nevertheless class dependent. At the dataset level, the learned mapping increases recall for *joy* from 90.98% under M_{hard} to 91.80%, while recall decreases from 67.40% to 65.88% for *surprise* and from 42.08% to 37.56% for *disgust*. These results indicate that probabilistic aggregation redistributes errors across coarse categories rather than uniformly improving top-1 classification.

The mechanism underlying these differences follows from how the fine-grained posterior is aggregated. Under the deterministic mapping, the complete fine-grained posterior is propagated through the one-hot transformation,

$$p^c_{\text{hard}}(x) = M_{\text{hard}} p^f(x),$$

so that the probability associated with each fine-grained emotion contributes entirely to its predefined coarse category. Probabilistic mapping instead uses

$$p^c_M(x) = M p^f(x),$$

where a column of M may distribute the probability associated with a fine-grained emotion across multiple coarse categories. Thus, all mapping strategies aggregate the same complete fine-grained posterior; they differ in whether each fine-grained category contributes deterministically or probabilistically across coarse categories.

Overall, the principal distinction between deterministic and probabilistic mapping is not a substantial improvement in aggregate top-1 classification. Rather, probabilistic mappings allow probability associated with each fine-grained category to be distributed across coarse categories instead of assigning it entirely to a single predefined parent, altering a small and class-dependent subset of coarse decisions. At the same time, the near identity of the argmax predictions produced by M_{emp} and M_θ shows that top-1 metrics alone provide limited information about the effect of learning the cross-granularity transformation.