Deferring Concept Bottleneck Models: Learning to Defer Interventions to Inaccurate Experts

Andrea Pugnana* University of Trento
Riccardo Massidda* University of Pisa
Francesco Giannini Scuola Normale Superiore
Pietro Barbiero IBM Research

Mateo Espinosa Zarlenga University of Cambridge, University of Oxford
Roberto Pellungrini Scuola Normale Superiore
Gabriele Dominici USI

Fosca Giannotti Scuola Normale Superiore
Davide Bacciu University of Pisa

Abstract

Concept Bottleneck Models (CBMs) are interpretable machine learning models that ground their predictions on human-understandable concepts, allowing for targeted interventions in their decision-making process. However, when intervened on, CBMs assume the availability of humans that can identify the need to intervene and always provide correct interventions. Both assumptions are unrealistic and impractical, considering labor costs and human error-proneness. In contrast, Learning to Defer (L2D) extends supervised learning by allowing machine learning models to identify cases where a human is more likely to be correct than the model, thus leading to deferring systems with improved performance. In this work, we gain inspiration from L2D and propose Deferring CBMs (DCBMs), a novel framework that allows CBMs to learn when an intervention is needed. To this end, we model DCBMs as a composition of deferring systems and derive a consistent L2D loss to train them. Moreover, by relying on a CBM architecture, DCBMs can explain the reasons for deferring on the final task. Our results show that DCBMs can achieve high predictive performance and interpretability by deferring only when needed.

1 Introduction

Concept Bottleneck Models (CBMs) [Koh et al., 2020] are a family of interpretable machine learning (ML) models that incorporate human-interpretable *concepts* as part of their training and predictive process. At test time, CBMs enable experts to correct any of their intermediate concepts' values, potentially triggering a change to the CBM's task prediction. This fosters a collaborative interaction between humans and AI systems, where a CBM may improve its accuracy when deployed with the support of an expert. However, CBMs suffer from a few shortcomings: first, increasing interpretability often comes at the expense of predictive accuracy, leading to an interpretability-accuracy trade-off [Zarlenga et al., 2022]; second, CBMs often assume that their set of concepts can fully predict the final task (i.e., they are *complete* [Yeh et al., 2020]); third, CBMs assume that human interventions are *infallible*, an over-simplification that does not reflect the real world where human experts may introduce errors, be unaware of their own potential weaknesses, and have a specific sub-expertise [Rastogi et al., 2022]. These practical limitations muddle the effects and dynamics of the human-AI collaboration expected when using CBMs.

*Equal Contribution (andrea.pugnana@unitn.it, riccardo.massidda@di.unipi.it).

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

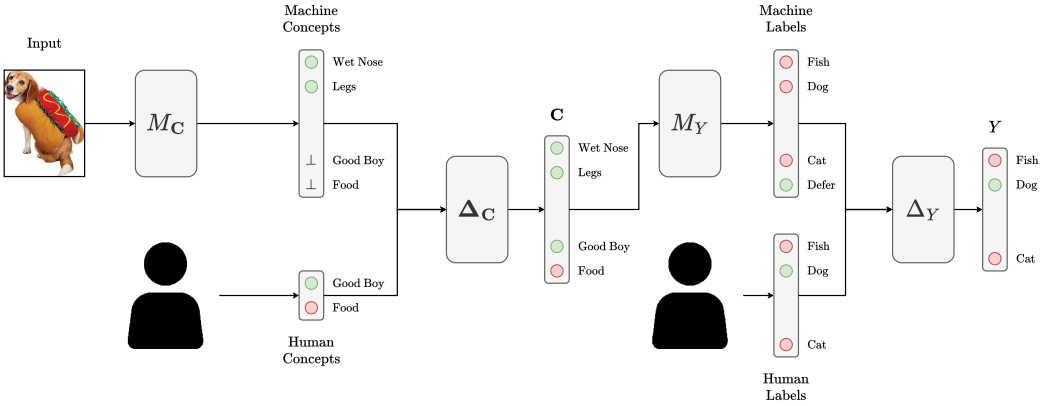


Figure 1: A DCBM: Given an input, the concept predictors $M_{\boldsymbol{C}}$ output either a concept's value or defer its prediction to a human (i.e., they predict $\perp$). Next, the deferring system $\Delta_{\boldsymbol{C}}$ outputs the human labels *only* on the deferred concepts, returning the system's predictions otherwise. The same applies to the final task, where the task classifier M_Y is an input of a dedicated deferring system Δ_Y. DCBMs can be trained by considering the cost of deferring, thus regulating the expected number of human deferrals.

To address the complex dynamics of human-in-the-loop interactions, Learning to Defer (L2D) has been introduced as an extension of supervised learning [Madras et al., 2018, Okati et al., 2021, Mozannar and Sontag, 2020]. In L2D, ML models can delegate challenging instances to human experts, enhancing human-AI team collaboration and outperforming both the ML models and the human experts [Mozannar et al., 2023]. Notably, conventional L2D approaches have been applied to single-classification tasks and are typically opaque, providing little insight into the reasons for deferring decisions [Ruggieri and Pugnana, 2025].

In this work, we introduce *Deferring Concept Bottleneck Models* (DCBMs), a novel class of models enabling learning to defer on CBMs (Figure 1). A key advantage of DCBMs is their ability to effectively learn when a concept or task prediction could benefit from human intervention. To the best of our knowledge, DCBM represents the first interpretable-by-design deferring system, enabling more transparent human-AI collaborations. Moreover, resorting to L2D, DCBMs introduce another variable to the accuracy-interpretability trade-off, i.e. the so-called *coverage*, which measures the percentage of times the ML model provides the prediction. Indeed, by allowing DCBMs to defer difficult cases to the human, one can achieve high accuracy and interpretability at the cost of deferring more to the human. Summarizing, our contributions are the following:

1. We introduce the Deferring Concept Bottleneck Model, a novel interpretable model capable of autonomously deferring on both its intermediate concepts and final task predictions (Section 3.1).
2. We propose a new deferral-aware loss for CBMs (Section 3.2) and prove that it is a consistent surrogate loss w.r.t. the intractable zero-one loss on the deferral procedure (Section 3.3).
3. We experimentally show how DCBMs react to varying costs and different human-accuracy degrees for defer (Section 4). Moreover, DCBMs can significantly improve concept-incomplete tasks.
4. Finally, we demonstrate how DCBMs can produce concept-based explanations for their final task deferrals by exploiting their interpretable-by-design architecture.

We organize the rest of the paper as follows. In Section 2, we introduce the background on CBMs and L2D. Then, in Section 3, we propose DCBMs and prove that their loss function is consistent with the L2D problem. Next, in Section 4, we report an empirical analysis highlighting the advantages of DCBMs. Finally, we discuss related works in Section 5 and summarize our work in Section 6.

2 Background

Given a variable V, we denote its domain as $\mathcal{D}(V)$, and its realization as $v \in \mathcal{D}(V)$. Similarly, we use bold for sets of variables $\boldsymbol{V}$ and their multi-variate realizations $\boldsymbol{v} \in \mathcal{D}(\boldsymbol{V})$.

Concept Bottleneck Models. Concept-based models are interpretable architectures that explain their predictions using high-level units of information known as "concepts" [Kim et al., 2018, Chen et al., 2020, Marconato et al., 2022, Kim et al., 2023, Barbiero et al., 2023, Oikarinen et al., 2023, Bortolotti et al., 2025]. Most of these approaches can be formulated as a Concept Bottleneck Model (CBM) [Koh et al., 2020], an architecture where predictions are made by composing (i) a *concept encoder* $g : \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{C})$ that maps samples $\boldsymbol{x} \in \mathcal{D}(\boldsymbol{X}) \subseteq \mathbb{R}^d$ (e.g., pixels) to a set of n_c concepts $\boldsymbol{c} \in \mathcal{D}(\boldsymbol{C}) = \{0, 1\}^{n_c}$ (e.g., "`red`", "`round`"), and (ii) a *task predictor* $f : \mathcal{D}(\boldsymbol{C}) \to \mathcal{D}(\boldsymbol{Y})$ that maps predicted concepts to a set of n_y tasks $\boldsymbol{y} \in \mathcal{D}(\boldsymbol{Y}) = \{0, 1\}^{n_y}$ (e.g., "`apple`", "`pear`").

CBMs can be trained (a) *independently*, where g and f are trained separately and later combined; (b) *sequentially*, where g is trained first, and its output is used to train f; or (c) *jointly*, where g and f are trained together. All of these training paradigms operate under the assumption that the training concept labels $\boldsymbol{c}$ are *complete*, meaning they are sufficient to predict the tasks $\boldsymbol{y}$ [Yeh et al., 2020].

Learning to Defer. Learning to Defer (L2D) [Madras et al., 2018] combines a human expert's knowledge, modeled as a *given* and non-trainable predictor $h \colon \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{Y})$, together with a *learnable* classifier $m \colon \mathcal{D}(\boldsymbol{X}) \to \mathcal{D}(\boldsymbol{Y}) \cup \{\bot\}$ over $|\boldsymbol{Y}| + 1$ classes, where the additional class, denoted as $\bot$, stands for the *deferral decision*. We define a deferring system as a pair $\Delta = (m, h)$ s.t.

$$\Delta(\boldsymbol{x}) = \begin{cases} m(\boldsymbol{x}) & \text{if } m(\boldsymbol{x}) \neq \bot \\ h(\boldsymbol{x}) & \text{otherwise.} \end{cases}$$

A deferring system is a human-AI team specifying who should predict between the human and the ML model. We stress here that the human predictions might *differ* from the ground-truth label, and thus trivially deferring each instance might not be optimal. According to Mozannar and Sontag [2020], L2D can be formalized as a risk minimization problem of the following zero-one loss:

$$\min_{m \in \mathcal{M}} \mathbb{E}_{\boldsymbol{x}, y, h \sim \mathbb{P}(\mathbf{x}, y, h)} \left[\mathbb{I}_{\{m(\boldsymbol{x}) \neq \bot\}} \mathbb{I}_{\{m(\boldsymbol{x}) \neq y\}} + \mathbb{I}_{\{m(\boldsymbol{x}) = \bot\}} \mathbb{I}_{\{h \neq y\}} \right] \tag{1}$$

where $\mathbb{P}(\mathbf{x}, y, h)$ is the distribution over (input, output, human-predictions) triplets, and $\mathcal{M}$ is the hypothesis spaces for the model m. Since directly optimizing Equation (1) is intractable, many *consistent surrogate losses*[2] have been proposed [Mozannar and Sontag, 2020, Verma and Nalisnick, 2022, Cao et al., 2023, Charusaie et al., 2022] to train single-task classifiers over the $|\boldsymbol{Y}| + 1$ classes. In a deferring system, the *coverage* counts the number of instances predicted by the model without deferring to a human, i.e., given a dataset $\left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{N}$, it corresponds to $1/N \cdot \sum_{i=1}^{N} \mathbb{I}_{\{m(\boldsymbol{x}^{(i)}) \neq \bot\}}$.

3 Deferring Concept Bottleneck Models

In this section, we first introduce Deferring Concept Bottleneck Models (DCBMs), a novel graphical probabilistic model for which we also define the exact but intractable learning-to-defer optimization problem (Section 3.1). Then, we introduce the surrogate loss and we show how it can be derived as the maximum likelihood of our graphical model (Section 3.2). Finally, we prove that our formulation results in a valid consistent loss for the L2D problem (Section 3.3), and we discuss how the loss' consistency can be ensured, while efficiently training DCBMs (Section 3.4).

3.1 Model Formulation

As in CBMs, we consider the problem of predicting both concept variables that directly depend on the input and task variables that are conditionally independent of the input given the concepts. We define DCBMs as an extension of CBMs, where each concept variable $C \in \boldsymbol{C}$ and each task variable $Y \in \boldsymbol{Y}$ is dealt with as a separate deferring system. Similar to CBMs, a DCBM can be framed as a probabilistic graphical model, with the difference that both concepts and tasks depend on human predictions when the model defers (Figure 2).

Let the set of concepts and task variables be $\boldsymbol{V} = \boldsymbol{C} \cup \boldsymbol{Y}$. We assign an expert H_V and a model $M_V \colon \mathcal{D}(\boldsymbol{Z}_V) \to [n_V] \cup \{\bot\}$ to each variable $V \in \boldsymbol{V}$. Here, the output space consists of $n_V + 1$ classes, including the deferral choice $\bot$, and either $\boldsymbol{Z}_V = \boldsymbol{X}$, if $V \in \boldsymbol{C}$, or $\boldsymbol{Z}_V = \boldsymbol{C}$, if

[2]Let ℓ and ℓ' be two loss functions. ℓ' is a consistent surrogate of ℓ whenever $\arg\min \ell' \subseteq \arg\min \ell$.

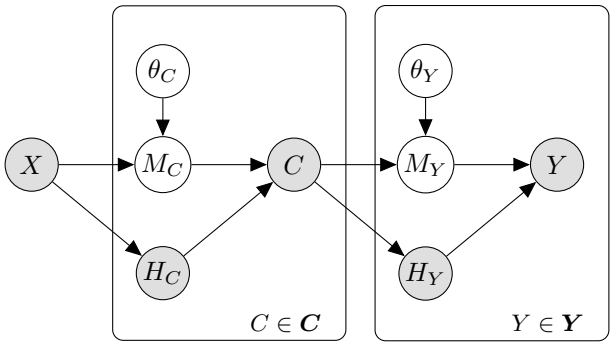


Figure 2: A DCBM is a Bayesian Network where inputs $\boldsymbol{X}$, concepts $\boldsymbol{C}$, tasks $\boldsymbol{Y}$, and human labels $\boldsymbol{H}$ are observed variables (in gray). As represented by the plate notation [Koller and Friedman, 2009], we assign a human expert and a latent model to each variable. We incorporate the deferral decision in the model through a dedicated output, denoted as $M = \bot$. Here, we learn each model M_V's parameters θ_V via maximum likelihood.

$V \in \boldsymbol{Y}$. Similarly, we denote the ground-truth output as k_V, which is to be intended as the label of either a concept or a task.

In contrast to traditional L2D setups, in a DCBM we need to train a model composed of several deferring systems, one for each concept and task variable. Hence, our objective would ideally minimize the number of mistakes made by all the deferring systems. This can be expressed through the following multi-variate zero-one loss, where we model the cost of each deferral via a hyperparameter $\lambda \in [0, 1]$:

Definition 1 (Multivariate Zero-One Loss)**.** *Given a set of variables $\boldsymbol{V}$ and a set of deferring systems $\boldsymbol{\Delta} = \{\Delta_V = (m_V, h_V)\}_{V \in \boldsymbol{V}}$ parameterized by a set of parameters $\boldsymbol{\theta} = \{\theta_V\}_{V \in \boldsymbol{V}}$, we define the multivariate zero-one loss as*

$$\sum_{V \in \boldsymbol{V}} \mathbb{I}_{\{m_V(\boldsymbol{z}_V) \neq \bot\}} \mathbb{I}_{\{m_V(\boldsymbol{z}_V) \neq k_V\}} + \mathbb{I}_{\{m_V(\boldsymbol{z}_V) = \bot\}} (\lambda + \mathbb{I}_{\{h_V \neq k_V\}}) \,, \tag{2}$$

where $\boldsymbol{z}_V$ and k_V are the realizations inputs and outputs, respectively, of each deferring system Δ_V.

3.2 Maximum Likelihood and Surrogate Loss

By deriving the negative log-likelihood from our probabilistic formulation of DCBMs (Figure 2), we can treat the maximum likelihood estimation of the parameters as a minimization problem. In this way, we obtain a loss function composed of two terms. Intuitively, the first term directly rewards the classifier for predicting the ground-truth class, while the second term rewards the model for deferring whenever the human prediction is correct.

Proposition 3.1 (Maximum Likelihood of DCBM)**.** *Let $\boldsymbol{\theta}$ be the parameters of a DCBM. Then, we can obtain the most likely parameters $\hat{\theta}$ given observations on the inputs $\boldsymbol{x}$, the concepts $\boldsymbol{c}$, the human $\boldsymbol{h}$, and the task $\boldsymbol{y}$, by minimizing the following loss function:*

$$\ell(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) = \sum_{V \in \boldsymbol{V}} \Big(\Psi\big(q(\boldsymbol{z}_V; \theta_V), k_V\big) + \mathbb{I}_{\{y_V = h_V\}} \Psi\big(q(\boldsymbol{z}_V; \theta_V), \bot\big) \Big) \tag{3}$$

where $q(\,\cdot\,; \theta_V) \colon \mathcal{D}(\boldsymbol{Z}_V) \to \mathbb{R}^{K_V + 1}$ returns the logits of the model M_V given $\boldsymbol{z}_V \in \mathcal{D}(\boldsymbol{Z}_V)$ and $\Psi(q(\boldsymbol{z}_V; \theta_V), k)$ is the negative log-probability of the class k given the logits $q(\boldsymbol{z}_V; \theta_V)$.

Proof. We report the proof in Appendix A.1. □

The negative log-likelihood we derived in Equation 3 does not take into account the cost of deferring. In this way, in scenarios where the human has a significant advantage, we can expect the model to underfit and almost always defer to the human [Mozannar et al., 2023]. To overcome these limitations, we define a penalized loss function by constraining the parameters of the model to enforce two additional conditions: (1) the model should not always defer when the human is correct, and (2) when

the human is not correct, the model should not defer. We report the formalization of the constrained optimization problem in Appendix A.2, and hereby report the resulting penalized loss,

$$\begin{aligned}\ell(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) = \sum_{V \in \boldsymbol{V}} & \Psi(q(\boldsymbol{z}_V; \theta_V), v) \\ & + (1-\lambda) \cdot \mathbb{I}_{\{y_V = h_V\}} \Psi(q(\boldsymbol{z}_V; \theta_V), \perp) \\ & + \lambda \cdot \mathbb{I}_{\{y_V \neq h_V\}} \sum_{k \in [K]} \Psi(q(\boldsymbol{z}_V; \theta_V), k),\end{aligned} \tag{4}$$

where $\lambda \in [0, 1]$ is an hyperparameter trading-off between deferrals and machine learning decisions.

In practice, the negative log-probability $\Psi(q(\boldsymbol{z}_V; \theta_V), k_V) = -\log P(M_V = k_V; \theta_V)$ of a class k_V according to the machine learning model M_V corresponds to the usual cross-entropy formulation

$$\Psi(q(\boldsymbol{z}), k) = -\log \left(\frac{\exp(q(\boldsymbol{z})_k)}{\sum_{k' \in [K+1]} \exp(q(\boldsymbol{z})_{k'})} \right). \tag{5}$$

The derivation of the negative log-likelihood (Equation 3) and its penalized counterpart (Equation 4) costitute an original contribution of this work. We further notice that in the univariate scenario our formulation collapses to known formulations from the L2D literature [Mozannar and Sontag, 2020, Eq. 6]. In this way, we first establish a clear connection between the maximum likelihood problem and the learning to defer task that, to the best of our knowledge, has not been previously identified in the literature. Further, we empirically consider also different formulations of the negative log-probability Ψ from the L2D literature — see Table 1 in Appendix A.2 for viable alternatives.

3.3 Loss Consistency

The multivariate scenario exacerbates the fact that always deferring to a human might not be the right choice, as the human's feedback may be incorrect or costly, and propagate such error. Therefore, to ensure that our model effectively defers to the human only when needed, we have to prove that the cost-free loss (Equation 3) and the penalized loss (Equation 4) of a DCBM are consistent surrogates of the ideal multivariate zero-one loss (Equation 2). We prove this by first showing that the sum of consistent losses on deferring systems with distinct parameters is consistent for the whole system.

Lemma 3.2. *Let $\ell'_1, \ell_1, \cdots, \ell'_m, \ell_m$ be possibly distinct loss functions. Assume that, for every $i \in \{1, \ldots, m\}$, $\ell'_i, \ell_i : \mathbb{R}^{n_i} \to \mathbb{R}$, being ℓ'_i a consistent surrogate of ℓ_i. Then $\ell' : \mathbb{R}^n \to \mathbb{R}$, with $n = n_1 + \ldots + n_m$ and $\ell'(\theta_1, \ldots, \theta_m) = \sum_{i=1}^m \ell'_i(\theta_i)$ is a consistent surrogate of $\ell : \mathbb{R}^n \to \mathbb{R}$, with $\ell(\theta_1, \ldots, \theta_m) = \sum_{i=1}^m \ell(\theta_i)$.*

Proof. We report the proof in Appendix A.3. □

Theorem 3.3. *The cost-free loss in Equation 3 and the DCBM penalized loss in Equation 4 are surrogate consistent losses of the multivariate zero-one loss of Equation 2 when $\boldsymbol{V} = \boldsymbol{C} \cup \boldsymbol{Y}$, and $\lambda = 0$ and $\lambda > 0$, respectively.*

Proof. We report the proof in Appendix A.4. □

Hence, under appropriate assumptions (whose practicalities we discuss in the next subsection), minimizing our novel surrogate losses corresponds to minimizing an exact multivariate zero-one loss.

3.4 Consistent Training of DCBMs

Theorem 3.3 ensures the consistency of our overall formulation, under a specific assumption on the loss functions being summed together: they should depend on disjoint sets of parameters. In essence, there are two main requirements to ensure consistency while training a DCBM. First, the model has to be trained *independently*, so that no information flows from the tasks' losses to the concepts' losses. Notably, independent training of CBMs is known to slightly decrease the performance compared to *jointly* training CBMs Koh et al. [2020]. However, independent training avoids the problem of concept leakage Mahinpei et al. [2021], Havasi et al. [2022], inherent to jointly trained models, thus maintaining the interpretability of the outcomes. For this reason, we focus on independently trained

models here and discuss additional experiments on jointly trained models, showing similar results to those seen for their independent counterparts, in Appendix E.

The second requirement concerns concept and task predictors, which should not share their parameters in the DCBM's architecture. Parameter sharing is common in CBMs, especially for computer vision tasks [Zarlenga et al., 2022], where an encoder produces a latent representation from the input space that is then fed to the concept predictors. To enable this in applications where parameter sharing is beneficial, we take the following two-step approach: first, we train an encoder to predict either all the concepts or the final task from the input features. Then, we freeze this encoder, discard the learned predictors, and independently train the concept predictors on the encoder's latent representation and the task predictor on the concepts using our consistent L2D loss. Still, for completeness' sake, we evaluate DCBMs when they share parameters across classifiers in Appendix E.

4 Experimental Evaluation

Our experimental analysis[3] aims to answer the following research questions:

- **Q1** Does deferring to a possibly imperfect human improve the performance of independently trained CBM-based approaches?
- **Q2** Does deferring mitigate the lack of completeness of a set of concepts for predicting a task?
- **Q3** Can DCBMs help to interpret *why* task classification was deferred?

4.1 Experimental Settings

Datasets. We perform our analysis on two real-world datasets: `cifar10-h` [Peterson et al., 2019] and `CUB` [Wah et al., 2011]. The `cifar10-h` dataset is a modified version of the `cifar10` dataset Krizhevsky et al. [2009] containing 10,000 images with both ground-truth and human-annotated labels. We adapted it for our scenario by adding as annotated concepts the 16 "superclass" concepts defined by Oikarinen et al. [2023] for each class. As human annotations are missing for the concepts, we treat humans as oracles on the concepts. Finally, `CUB` is a dataset commonly used for image classification with CBMs. We consider the complete set of 112 concepts used by Koh et al. [2020]. Since the dataset reports annotator uncertainty on the concepts, we use them to produce random human concept labels as done by Collins et al. [2023] (Appendix B). In the `CUB` task label, however, we treat humans as oracles. Finally, we employ the synthetic `completeness` [Laguna et al., 2024] dataset to study possible variants of our method, whose results we report in Appendix E.

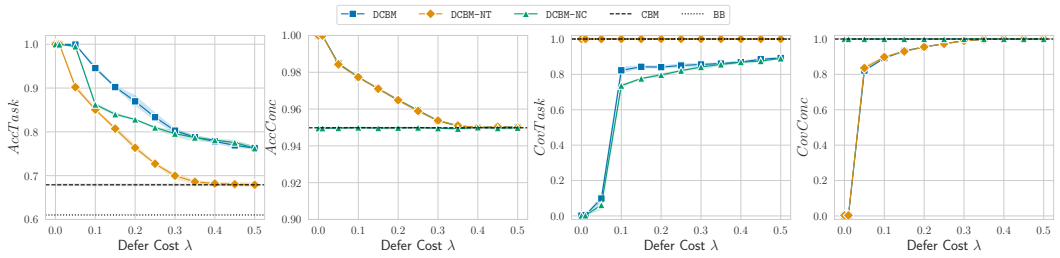
Methods. We compare a complete DCBM architecture (`DCBM`) with some ablated variants and baselines. In particular, we consider the following baselines: (i) a black-box model trained with standard supervised learning on the final task only (`BB`) (ii) a standard CBM without the deferring option (`CBM`). To evaluate the effect of deferring on concepts and tasks, we also compare with the following ablations: a DCBM that can not defer on the final task (`DCBM-NT`) and a DCBM that can not defer on the concepts (`DCBM-NC`). In all the datasets, we train the models using the state-of-the-art Asymmetric Softmax [Cao et al., 2023, ASM] parameterization of the negative-log-likelihood in our loss functions. We provide further details on the adopted architectures and the experimental setup in Appendix B. We report in Appendix E results for other losses on the `completeness` dataset.

Metrics. We report four main metrics: the accuracy on the final task (`AccTask`), the average accuracy among all concepts (`AccConc`), the coverage of the model on the final task (`CovTask`), which counts how many times the model directly classified the task label, and the average coverage on all concepts (`CovConc`), which is the percentage of concepts that are not deferred.

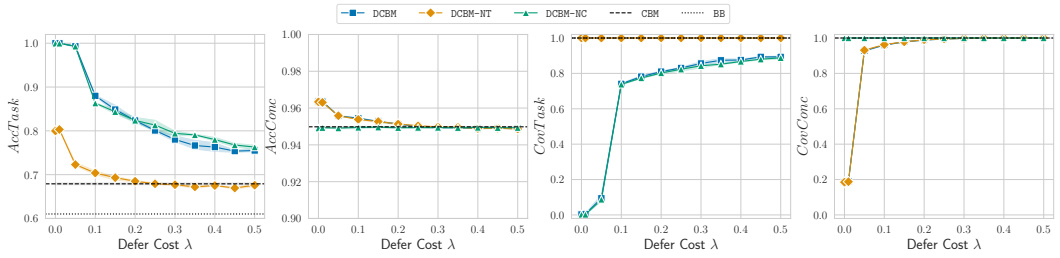
4.2 Experimental Results

Q1: Improving CBM's Performance with Deferring. We study the performance improvement on CBMs on the `CUB` dataset. First, we consider the ideal scenario where the human predictions match the ground-truth (Figure 3a). Then, we exploit the human uncertainty annotations on the `CUB` dataset to study a scenario where the human might wrongly classify a concept (Figure 3b).

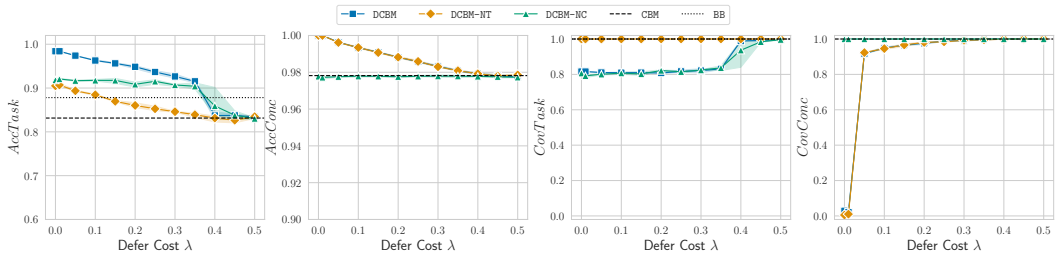
[3]We provide the code for reproducing our experiments at https://github.com/andrepugni/DCBM.



(a) CUB: Oracle Human on Concepts



(b) CUB: Uncertain Human Annotations on Concepts



(c) cifar10-h: Oracle Human on Concepts

Figure 3: Results on CUB dataset when human experts have perfect accuracy on the concepts (Figure 3a); on CUB dataset when human experts do not have perfect accuracy on the concepts (Figure 3b); on cifar10-h dataset when human experts have perfect accuracy on the concepts but not on the final task (Figure 3c) We report each metric's average and standard deviation as we increase the defer costs λ. The CBM and BB baselines are constant as they are independent of the defer cost. DCBM outperform competing baselines for lower deferral costs λ. Increasing the cost λ reduces the number of deferrals and decreases the DCBM performance *up to* the standard CBM performance.

In the first scenario, when the defer costs are significantly low ($\lambda < 0.05$), the deferring systems tend to over-rely on the human and thus the coverage of the machine learning model is zero for both concepts and tasks. As expected, increasing the defer cost increases both the coverage on the task (CovTask) and on the concepts (CovConc). At the same time, it also reduces the accuracy of the prediction, which is, however, still *over* the standard CBM baseline without deferring capabilities. In summary, the performance of the ablated (DCBM-NC, DCBM-NT) and the full model (DCBM) tend to those of the standard non-deferring CBM for higher defer costs, while improving performance for lower defer costs. Therefore, as a standard practice in the L2D literature [Wei et al., 2024], by leveraging the defer cost, we can ensure that the deferring systems do not over-rely on the human.

Notably, in the scenario where humans might wrongly classify some of the concept labels (Figure 3b), the results emphasize how the human expert's ability to correctly predict concepts affects the DCBM. In the presence of potentially incorrect humans on the concepts, the model correctly learns when there is no advantage in deferring concept predictions to humans. In detail, when the defer cost is zero ($\lambda = 0$), DCBM still has a non-zero coverage (CovConc ≈ 0.2), meaning that it is not deferring 20% of instances to humans, even if it would be "free" to do so. We provide additional results in Appendix E, showing how competitive intervention strategies fail to capture this aspect.

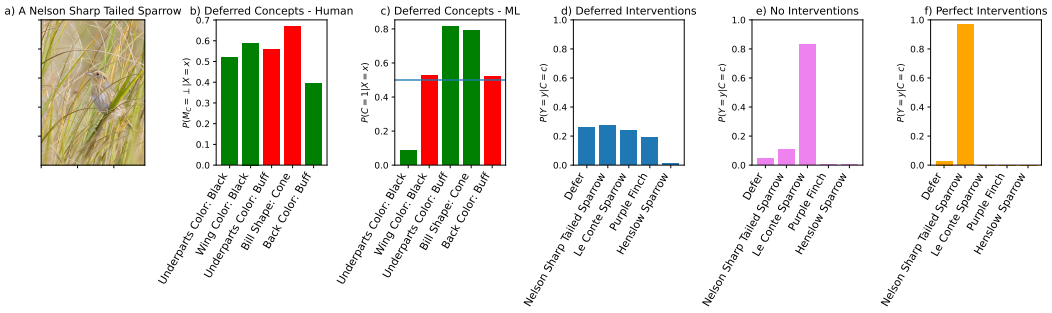


Figure 4: Interpretation of a DCBM with defer cost $\lambda = 0.1$ on an input sample. From left to right: (a) an example of an image from the CUB dataset; (b) the concepts that the model has deferred with the estimated probability, green bars stand for when the human correctly predicts the concept, red otherwise; (c) the estimated probability of each deferred concept being true according to the machine learning model, green bars stand for when the ML would have correctly predicted the concept, red otherwise; (d) the estimated probability of top-5 final task labels after deferring the concepts to the human (standard DCBM behavior); (e) the estimated probability of top-5 final task labels without deferring the concepts to the human; (f) the estimated probability of top-5 final task labels from the ground-truth concepts.

Therefore, the DCBMs automatically adjust the coverage depending not only on the defer cost but also on the human competence, extending traditional CBMs to account also for incorrect humans.

Q2: Addressing Incompleteness through Human-AI collaboration. We investigate the impact of deferring in an incomplete scenario, where the set of concepts is not sufficient to distinguish between two or more classes. In practice, this means that we cannot learn a good task classifier for those instances with the same concept-level representation — such as `cat` and `deer` in the `cifar10-h` dataset. A better choice would then be to defer to a human, who can distinguish between the two classes by also employing input data or additional information.

We validate our hypothesis on `cifar10-h` (see Figure 3c): results show that for low defer costs, the DCBM outperforms other baselines, with an $\texttt{AccTask} \approx 0.984$ at $\lambda = 0$ and a $\texttt{CovTask} \approx 0.816$. Furthermore, deferring on both the concepts and the task proves to be better than deferring only on one of the two, as shown by the results of our ablated models DCBM-NC and DCBM-NT.

Since on `cifar10-h` the human expert can make mistakes on the final task, our DCBM model correctly identifies that always deferring to the human would not be optimal: even when deferring to a human would be "free" ($\lambda = 0.0$), the DCBM has a high coverage on the classification task ($\texttt{CovTask} \approx 0.8$). The previously discussed raise in performance then comes from the DCBM correctly deferring to a human *only* when beneficial. Indeed, as shown in Appendix E, deferral occurs only for `cat` and `deer` instances. Moreover, increasing the defer cost λ decreases the classification performance while increasing the coverage of the machine learning model. However, it is worth remarking that our DCBM still performs better than the CBM baseline even for higher costs ($\lambda \approx 0.3$), where no concepts are deferred to the human ($\texttt{CovConc} \approx 1.0$) and most task classifications are performed by the machine ($\texttt{CovTask} \approx 0.8$).

In a nutshell, DCBMs provide a useful mechanism to deploy a model in an incomplete setting, addressing the risks that might arise from deploying a classifier that arbitrarily chooses one of the two (or possibly more) entangled classes.

Q3: Interpretable Learning to Defer in DCBMs. We show how DCBMs can help interpret the reasons for the final task deferring, by following a similar approach to Zarlenga et al. [2023]. For this purpose, we consider the CUB dataset and a DCBM trained with cost $\lambda = 0.1$. We discuss here an instance of bird image to be classified (Figure 5) and report further examples in Appendix C.

A main advantage of DCBMs is their capability to identify which concepts should be corrected by a human intervention, without additional interventions. Since the human supervision is not perfectly accurate in the CUB dataset, the interventions might not correspond to the ground-truth values (Figure 4b). Interestingly, the deferred concepts are not easy to grasp in the original image:

the underparts are partially covered by grass blades, the bill is not clearly cone-shaped, and the back is not really visible in the image. In general, highlighting the particular concepts on which the human should be a safer option than the ML model favours the interpretation of the classifier. Moreover, we also stress that without deferring to the human, the machine learning model would have wrongly classified some deferred concepts (Figure 4c).

This example also stresses that DCBMs try to defer only when worthy and necessary, without involving humans when they are likely to make mistakes. Moreover, interventions allows us to reason on how concepts effectively lead the DCBM to predict the correct class (Figure 4d). In fact, without intervening on the concepts, the final task model would have predicted a wrong label (Figure 4e). Finally, we also report how the final task model would have behaved under perfect interventions, i.e., where a human has access to ground-truth concept labels, as assumed by the standard CBM literature (Figure 4f). As expected, the DCBM would increase its own confidence in the correct label, at the cost of human supervision on *all* of the 112 concepts instead of the *only five* identified by the DCBM.

5 Related Works

Deferring systems. L2D, as introduced in Madras et al. [2018], is an instance of hybrid decision-making where humans oversee machines. Since directly optimizing Equation 1 is NP-hard even in simple settings [Mozannar et al., 2023], Mozannar and Sontag [2020] proposed consistent surrogate losses, which have since become the standard approach for jointly learning the deferral policy and the ML predictor [Charusaie et al., 2022, Verma and Nalisnick, 2022, Mozannar et al., 2023, Cao et al., 2023, Liu et al., 2024, Wei et al., 2024]. A formal characterization of humans in the loop is provided by Okati et al. [2021]. Recent works extend the L2D problem to account for multiple human experts, e.g., see Verma et al. [2023], Mao et al. [2023] and Cao et al. [2023], cases where the ML model is already given and not jointly trained, e.g., [Charusaie et al., 2022, Mao et al., 2023, Montreuil et al., 2025a,b], and how they relate to causal frameworks [Palomba et al., 2025, Gao and Yin, 2025].

Concept Interventions. CBMs have seen a growth of interest in the context of *concept interventions*, operations that improve a CBM's overall task performance in the presence of test-time human feedback. Works in this area have explored (1) how to best select which concepts to intervene on next when interventions are costly [Shin et al., 2023, Chauhan et al., 2023] — see Appendix D for a comparison between DCBMs and Uncertainty on Concept prediction [UCP; Shin et al. 2023], a popular and competitive strategy to perform interventions — (2) how to improve a model's receptiveness to interventions and learn an *intervention policy* [Zarlenga et al., 2023], and (3) how to intervene on otherwise black-box models [Laguna et al., 2024]. In particular, while policy-based methods [Chauhan et al., 2023] rank which interventions should be prioritized to enhance the classification performance of the model, they still require the human expert to initiate the procedure to request an intervention. By modeling the human predictive performance as done by the L2D literature, DCBMs ask instead for interventions without further supervision at inference time. Other approaches have exploited inter-concept relationships to propagate single-concept interventions [Vandenhirtz et al., 2024, Raman et al., 2024, Dominici et al., 2025] and have used interventions as sources of continual learning labels [Steinmann et al., 2024]. Finally, Sheth and Kahou [2023] and Collins et al. [2023] both discuss notions of supervisor uncertainty, where we may be interested in modeling errors from an expert performing interventions. Nevertheless, works on concept interventions fundamentally differ from our L2D-based approach in that they assume that experts themselves trigger a correction in a model's concept predictions. This makes it difficult for these approaches to adapt to expert-specific competencies and to be easily deployed in practice where it is desirable to know *when* a human should be called to intervene.

6 Conclusions and Future Work

This paper introduces DCBM, a novel approach that allows CBMs to defer to a human without additional supervision. By training the CBM with an especially designed learning to defer loss function, a DCBM can implicitly model the predictive distribution of the human, and thus defer only on instances where the expert is more likely to be correct of the machine learning model. Moreover, we formally proved the consistency of our deferring loss function for independent training of CBMs. Our experimental results highlight that DCBMs effectively learn when to involve a human, boosting

overall predictive performance only when the human is better than the ML model. Moreover, directly involving a human helps mitigate cases where concepts are incomplete. Finally, the interpretable by-design nature of DCBMs offers ways to audit the deferring systems, showing promising results in explaining their limits.

Limitations and Future Works. We acknowledge a few limitations of our current work. First, the actual implementations of CBMs and deferring systems in real-life settings is still overlooked [Ruggieri and Pugnana, 2025]. Hence, user studies are highly needed to evaluate how humans can benefit from ours and other concept-based approaches.

Second, our theoretical approach considers concepts as independent variables. While DCBMs can be directly extended to group together sets of mutually exclusive binary concepts (thus not independent) into a single multi-class concept, we are not considering more complex relationships among concepts. Extending our approach to account for a hierarchical structure of concepts — as done e.g., in causal abstraction [Geiger et al., 2021, Massidda et al., 2024] — is also an open research direction.

Third, in this paper, we consider a single expert per concept/task and implicitly assume that the expert's costs are the same when deferring at the concept and task level. While our framework can be easily extended to account for different costs and multi-expert settings [Mao et al., 2023], we did not explicitly investigate how these modelling choices can lead to different DCBMs. Furthermore, as all L2D approaches, our proposal models the predictive distribution of the human expert and assumes that the human will follow the same distribution at test time. Changes in how the human tackles the same task might affect the performance of the overall human-AI system. Detecting such distribution shifts and integrating with continual learning strategies [Parisi et al., 2019] could then help real-world applications and constitutes a promising research line for L2D methods in general.

Finally, regarding the explanation of deferring systems, CBMs are interpretable models that can be used to enlighten the decision process toward either a class or a defer prediction. This paper used defer on concepts and interventions to provide explanations on task's deferral. Since several methodologies for explaining CBMs have been considered, such as those based on logical rules, DCBMs could be extended to account for different kinds of explanations.

Acknowledgments and Disclosure of Funding

We thank Martina Cinquini for initial discussion and feedback on the article.
This work has been funded by the European Union under Grant Agreement no. 101120763 - TANGO. This work has been supported by the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI" and ERC-2018-ADG G.A. 834756 "XAI: Science and technology for the eXplanation of AI decision making". This work has been supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No 952215. This work has been partially supported by IMAGINE, a project funded by the Swiss National Science Foundation (No. 224226). GD acknowledges support from the European Union's Horizon Europe project SmartCHANGE (No. 101080965). MEZ acknowledges that the majority of this work was done with the support of the Gates Cambridge Trust via a Gates Cambridge Scholarship.

References

Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Mateo Espinosa Zarlenga, Lucie Charlotte Magister, Alberto Tonda, Pietro Lio, Frédéric Precioso, Mateja Jamnik, and Giuseppe Marra. Interpretable neural-symbolic concept reasoning. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 1801–1825. PMLR, 2023. 3

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In *NeurIPS*, 2025. 3

Yuzhou Cao, Hussein Mozannar, Lei Feng, Hongxin Wei, and Bo An. In defense of softmax parametrization for calibrated and consistent learning to defer. In *NeurIPS*, 2023. 3, 6, 9, 19

Mohammad-Amin Charusaie, Hussein Mozannar, David A. Sontag, and Samira Samadi. Sample efficient learning of predictors that complement humans. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 2972–3005. PMLR, 2022. 3, 9

Kushal Chauhan, Rishabh Tiwari, Jan Freyberg, Pradeep Shenoy, and Krishnamurthy Dvijotham. Interactive concept bottleneck models. In *AAAI*, pages 5948–5955. AAAI Press, 2023. 9

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 3

Katherine Maeve Collins, Matthew Barker, Mateo Espinosa Zarlenga, Naveen Raman, Umang Bhatt, Mateja Jamnik, Ilia Sucholutsky, Adrian Weller, and Krishnamurthy Dvijotham. Human uncertainty in concept-based AI systems. In *AIES*, pages 869–889. ACM, 2023. 6, 9

Gabriele Dominici, Pietro Barbiero, Mateo Espinosa Zarlenga, Alberto Termine, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. Causal concept graph models: Beyond causal opacity in deep learning. In *ICLR*. OpenReview.net, 2025. 9

Ruijiang Gao and Mingzhang Yin. Confounding-robust deferral policy learning. In *AAAI*, pages 14238–14246. AAAI Press, 2025. 9

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *NeurIPS*, pages 9574–9586, 2021. 10

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *NeurIPS*, 2022. 5

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 20

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2673–2682. PMLR, 2018. 3

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 16521–16540. PMLR, 2023. 3

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 21

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 2020. 1, 3, 5, 6, 21

Daphne Koller and Nir Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009. 4

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Learning multiple layers of features from tiny images. Master's thesis, University of Toronto, 2009. URL `http://www.cs.toronto.edu/~kriz/cifar.html`. 6, 20

Sonia Laguna, Ricards Marcinkevics, Moritz Vandenhirtz, and Julia E. Vogt. Beyond concept bottleneck models: How to make black boxes intervenable? In *NeurIPS*, 2024. 6, 9, 24

Shuqi Liu, Yuzhou Cao, Qiaozhen Zhang, Lei Feng, and Bo An. Mitigating underfitting in learning to defer with consistent losses. In *AISTATS*, volume 238 of *Proceedings of Machine Learning Research*, pages 4816–4824. PMLR, 2024. 9, 18, 19, 20, 25

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019. 21

David Madras, Toniann Pitassi, and Richard S. Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *NeurIPS*, pages 6150–6160, 2018. 2, 3, 9

Anita Mahinpei, Justin Clark, Isaac Lage, Finale Doshi-Velez, and Weiwei Pan. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021. 5

Anqi Mao, Christopher Mohri, Mehryar Mohri, and Yutao Zhong. Two-stage learning to defer with multiple experts. In *NeurIPS*, 2023. 9, 10

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretable, leak-proof concept-based models. In *NeurIPS*, 2022. 3

Riccardo Massidda, Sara Magliacane, and Davide Bacciu. Learning causal abstractions of linear structural causal models. In *UAI*, volume 244 of *Proceedings of Machine Learning Research*, pages 2486–2515. PMLR, 2024. 10

Yannis Montreuil, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Adversarial Robustness in Two-Stage Learning-to-Defer: Algorithms and Guarantees. In *ICML*. OpenReview.net, 2025a. 9

Yannis Montreuil, Yeo Shu Heng, Axel Carlier, Lai Xing Ng, and Wei Tsang Ooi. Two-stage learning-to-defer for multi-task learning. In *ICML*. OpenReview.net, 2025b. 9

Hussein Mozannar and David A. Sontag. Consistent estimators for learning to defer to an expert. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 7076–7087. PMLR, 2020. 2, 3, 5, 9, 20

Hussein Mozannar, Hunter Lang, Dennis Wei, Prasanna Sattigeri, Subhro Das, and David A. Sontag. Who should predict? exact algorithms for learning to defer to humans. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pages 10520–10545. PMLR, 2023. 2, 4, 9, 19

Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *ICLR*. OpenReview.net, 2023. 3, 6

Nastaran Okati, Abir De, and Manuel Gomez-Rodriguez. Differentiable learning under triage. In *NeurIPS*, pages 9140–9151, 2021. 2, 9

Filippo Palomba, Andrea Pugnana, José M. Álvarez, and Salvatore Ruggieri. A causal framework for evaluating deferring systems. In *AISTATS*, volume 258 of *Proceedings of Machine Learning Research*, pages 2143–2151. PMLR, 2025. 9, 20

German I Parisi, Ronald Kemker, Jose L Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural networks*, 113:54–71, 2019. 10

Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and Olga Russakovsky. Human uncertainty makes classification more robust. In *ICCV*, pages 9616–9625. IEEE, 2019. 6

Naveen Raman, Mateo Espinosa Zarlenga, and Mateja Jamnik. Understanding inter-concept relationships in concept-based models. In *ICML*. OpenReview.net, 2024. 9

Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R. Varshney, Amit Dhurandhar, and Richard Tomsett. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proc. ACM Hum. Comput. Interact.*, 6(CSCW1):83:1–83:22, 2022. 1

Salvatore Ruggieri and Andrea Pugnana. Things machine learning models know that they don't know. In *AAAI*, pages 28684–28693. AAAI Press, 2025. 2, 10

Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. In *NeurIPS*, 2023. 9

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pages 31504–31520. PMLR, 2023. 9, 23

David Steinmann, Wolfgang Stammer, Felix Friedrich, and Kristian Kersting. Learning to intervene on concept bottlenecks. In *ICML*. OpenReview.net, 2024. 9

Moritz Vandenhirtz, Sonia Laguna, Ricards Marcinkevics, and Julia E. Vogt. Stochastic concept bottleneck models. In *NeurIPS*, 2024. 9

Rajeev Verma and Eric T. Nalisnick. Calibrated learning to defer with one-vs-all classifiers. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 22184–22202. PMLR, 2022. 3, 9

Rajeev Verma, Daniel Barrejón, and Eric T. Nalisnick. Learning to defer to multiple experts: Consistent surrogate losses, confidence calibration, and conformal ensembles. In *AISTATS*, volume 206 of *Proceedings of Machine Learning Research*, pages 11415–11434. PMLR, 2023. 9, 19

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 6

Zixi Wei, Yuzhou Cao, and Lei Feng. Exploiting human-ai dependence for learning to defer. In *ICML*. OpenReview.net, 2024. 7, 9

Chih-Kuan Yeh, Been Kim, Sercan Ömer Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *NeurIPS*, 2020. 1, 3

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*. BMVA Press, 2016. 20

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frédéric Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, 2022. 1, 6, 21

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. In *NeurIPS*, 2023. 8, 9

Supplementary Material

Table of Contents

A Proofs

A.1 Proposition 3.1 — Maximum Likelihood of DCBM

In this proof, we report the derivation of the maximum likelihood of the Bayesian Network corresponding to our Deferring Concept Bottleneck Model (DCBM), which we reported in Figure 2. We assume that our dataset is composed of i.i.d. samples from the joint distribution of the observable variables. Therefore, we consider the input data $\boldsymbol{x} \in \mathcal{D}(\boldsymbol{X})$, the concept values $\boldsymbol{c} \in \mathcal{D}(\boldsymbol{C})$, the task values $\boldsymbol{y} \in \mathcal{D}(\boldsymbol{Y})$, and the human annotations on both concepts and tasks $\boldsymbol{h} \in \mathcal{D}(\boldsymbol{H})$. We first define the likelihood of the data by marginalizing over the latent variables, i.e., of each concept model M_C and task model M_Y for all variables $C \in \boldsymbol{C}$ and $Y \in \boldsymbol{Y}$. We recall that for each concept model M_C we have one out of $n_C + 1$ possible outcomes, where n_C is the number of possible realizations of C and the additional value accounts for the deferred decision, as in $M_C = \perp$. Similarly, each task model M_Y has $n_Y + 1$ possible outcomes. We then marginalize one variable at a time from the joint likelihood, starting from an arbitrary task variable $Y \in \boldsymbol{Y}$.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) = p(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h} \mid \boldsymbol{\theta}) \tag{6}$$

$$= \sum_{k \in [n_Y+1]} p(\boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}, M_Y = k \mid \boldsymbol{\theta}) \tag{7}$$

$$= p(\boldsymbol{x}) \sum_{k \in [n_Y+1]} p(\boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}, M_Y = k \mid \boldsymbol{x}, \boldsymbol{\theta}) \tag{8}$$

$$= p(\boldsymbol{x}) \sum_{k \in [n_Y+1]} p(Y = y \mid M_Y = k, H_Y = h_Y) p(M_Y = k \mid \boldsymbol{c}, \theta_Y) p(H_Y = h_Y \mid \boldsymbol{x}, \boldsymbol{c}) \cdot \; p(\boldsymbol{c}, \boldsymbol{y}_{\setminus Y}, \boldsymbol{h}_{\setminus Y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\setminus \theta_Y}), \tag{9}$$

$$= p(\boldsymbol{x}) p(H_Y = h_Y \mid \boldsymbol{x}, \boldsymbol{c}) \sum_{k \in [n_Y+1]} p(Y = y \mid M_Y = k, H_Y = h_Y) p(M_Y = k \mid \boldsymbol{c}, \theta_Y) \cdot \; p(\boldsymbol{c}, \boldsymbol{y}_{\setminus Y}, \boldsymbol{h}_{\setminus Y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\setminus \theta_Y}), \tag{10}$$

where we employ the operator $\setminus$ to denote the removal from a set of a variable.

Before marginalizing the remaining variables, we focus on the sum over possible values of the model M_Y. We can further decompose it by considering whether the model value is a possible value for Y or a deferral $\perp$. According to our definition of a deferring system, $Y = y$ if and only if the model has value $M_Y = y$ or it deferred the decision through $M_Y = \perp$ but for the human expert holds $H_Y = y$. Therefore, it holds that $p(Y = y \mid M_Y = k, H_Y = h_y) = 1$ if and only if $M_Y = y$ whenever $M_Y \neq \perp$ and $p(Y = y \mid M_Y = \perp, H_Y = h_y) = 1$ if and only if $h_Y = y$ whenever $M_Y = \perp$. Formally,

$$\sum_{k \in [n_Y+1]} p(Y = y \mid M_Y = k, h_Y) p(M_Y = k \mid \boldsymbol{c}, \theta_Y) \tag{11}$$

$$= \sum_{k \in [n_Y]} p(Y = y \mid M_Y = k, h_Y) p(M_Y = k \mid \boldsymbol{c}, \theta_Y) + p(Y = y \mid M_Y = \perp, h_Y) p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y) \tag{12}$$

$$= \sum_{k \in [n_Y]} \mathbb{I}[y = k] p(M_Y = k \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y] p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y) \tag{13}$$

$$= p(M_Y = y \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y] p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y) \tag{14}$$

where $\mathbb{I}[\cdot]$ is the indicator function taking value one if the proposition is true, zero otherwise.

Therefore, we can apply the same decomposition to all tasks $\boldsymbol{Y}$ and rearrange terms as follows.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) \tag{15}$$

$$= p(\boldsymbol{x})p(h_Y \mid \boldsymbol{x}, \boldsymbol{c})\big(p(M_Y = y \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y]p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y)\big)p(\boldsymbol{c}, \boldsymbol{y}_{\setminus Y}, \boldsymbol{h}_{\setminus Y} \mid \boldsymbol{x}, \boldsymbol{\theta}_{\setminus \theta_Y}) \tag{16}$$

$$= p(\boldsymbol{x})p(h_Y \mid \boldsymbol{x}, \boldsymbol{c}) \prod_{Y \in \boldsymbol{Y}} \big(p(M_Y = y \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y]p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y)\big)p(\boldsymbol{c}, \boldsymbol{h_C} \mid \boldsymbol{x}, \boldsymbol{\theta_C}) \tag{17}$$

$$= p(\boldsymbol{x})p(h_Y \mid \boldsymbol{x}, \boldsymbol{c})p(\boldsymbol{c}, \boldsymbol{h_C} \mid \boldsymbol{x}, \boldsymbol{\theta_C}) \prod_{Y \in \boldsymbol{Y}} \big(p(M_Y = y \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y]p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y)\big). \tag{18}$$

Then, we can apply a similar decomposition to concepts, starting for an arbitrary concept $C \in \boldsymbol{C}$.

$$p(\boldsymbol{c}, \boldsymbol{h_C} \mid \boldsymbol{x}, \theta_{\boldsymbol{C}}) \tag{19}$$

$$= \sum_{k \in [n_C + 1]} p(\boldsymbol{c}, \boldsymbol{h_C}, M_C = c \mid \boldsymbol{x}, \theta_{\boldsymbol{C}}) \tag{20}$$

$$= p(h_C \mid \boldsymbol{x}) \sum_{k \in [n_C + 1]} p(C = c \mid M_C = k, h_C)P(M_C = c \mid \boldsymbol{x}, \theta_C) \cdot p(\boldsymbol{c}_{\setminus C}, h_{\boldsymbol{C} \setminus C} \mid \boldsymbol{x}, \theta_{\boldsymbol{C} \setminus C}) \tag{21}$$

$$= p(h_C \mid \boldsymbol{x})\big(p(M_C = c \mid \boldsymbol{x}, \theta_C) + \mathbb{I}[h_C = c]p(M_C = \perp \mid \boldsymbol{x}, \theta_C).\big) \cdot p(\boldsymbol{c}_{\setminus C}, h_{\boldsymbol{C} \setminus C} \mid \boldsymbol{x}, \theta_{\boldsymbol{C} \setminus C}) \tag{22}$$

$$= p(h_{\boldsymbol{C}} \mid \boldsymbol{x}) \prod_{C \in \boldsymbol{C}} \big(p(M_C = c \mid \boldsymbol{x}, \theta_C) + \mathbb{I}[h_C = c]p(M_C = \perp \mid \boldsymbol{x}, \theta_C)\big). \tag{23}$$

Finally, leading to the following form, which we further simplify by denoting as $\boldsymbol{z}_V$ the input of each variable $V \in \boldsymbol{V}$ and as $v \in \mathcal{D}(V)$ its realization in the dataset.

$$\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) \tag{24}$$

$$\begin{aligned} = p(\boldsymbol{x})p(\boldsymbol{h_C} \mid \boldsymbol{x})p(\boldsymbol{h_Y} \mid \boldsymbol{x}, \boldsymbol{c}) &\prod_{C \in \boldsymbol{C}} (p(M_C = c \mid \boldsymbol{x}, \theta_C) + \mathbb{I}[h_C = c]p(M_C = \perp \mid \boldsymbol{x}, \theta_C)) \\ &\cdot \prod_{Y \in \boldsymbol{Y}} (p(M_Y = v \mid \boldsymbol{c}, \theta_Y) + \mathbb{I}[h_Y = y]p(M_Y = \perp \mid \boldsymbol{c}, \theta_Y).) \end{aligned} \tag{25}$$

$$= p(\boldsymbol{x})p(\boldsymbol{h} \mid \boldsymbol{x}, \boldsymbol{c}) \prod_{V \in \boldsymbol{V}} (p(M_V = v \mid \boldsymbol{z}_V, \theta) + \mathbb{I}[h_V = v]p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V).) \tag{26}$$

Finally, we show that maximizing the likelihood equates to minimizing the loss function we defined in Section 3. We recall that to this end, we assume to have for each variable $V \in \boldsymbol{V}$ a machine learning model $g(\cdot; \theta_V)$ that produces $n_V + 1$ activations, one for each class and one additional for the defer action.

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{c}, \boldsymbol{y}, \boldsymbol{h}) \tag{27}$$

$$= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{x})p(\boldsymbol{h} \mid \boldsymbol{x}, \boldsymbol{c}) \prod_{V \in \boldsymbol{V}} p(M_V = v \mid \boldsymbol{z}_V, \theta) + \mathbb{I}[h_V = v]p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V) \tag{28}$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{V \in \boldsymbol{V}} p(M_V = v \mid \boldsymbol{z}_V, \theta) + \mathbb{I}[h_V = v]p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V) \tag{29}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{V \in \boldsymbol{V}} \log(p(M_V = v \mid \boldsymbol{z}_V, \theta) + \mathbb{I}[h_V = v]p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V)) \tag{30}$$

$$= \arg\max_{\boldsymbol{\theta}} \sum_{V \in \boldsymbol{V}} \log(p(M_V = v \mid \boldsymbol{z}_V, \theta)) + \mathbb{I}[h_V = v] \log(p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V)) \tag{31}$$

$$= \arg\min_{\boldsymbol{\theta}} \sum_{V \in \boldsymbol{V}} -\log(p(M_V = v \mid \boldsymbol{z}_V, \theta)) - \mathbb{I}[h_V = v] \log(p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V)) \tag{32}$$

$$= \arg\min_{\boldsymbol{\theta}} \sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + \mathbb{I}[h_V = v]\Psi(q(\boldsymbol{z}_V; \theta_V), \perp), \tag{33}$$

where $\Psi(g(\boldsymbol{z}_V; \theta_V))$ then corresponds to the standard formulation with the softmax operator, reported in Table 1. In the same table, we report alternative formulations for the same object from the learning to defer literature. Further, we can justify the transition from Equation (30) to Equation (31) since the following holds

$$\begin{aligned} &\log(p(M_V = v \mid \boldsymbol{z}_V, \theta) + \mathbb{I}[h_V = v] p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V)) \\ \geq &\log(p(M_V = v \mid \boldsymbol{z}_V, \theta)) + \mathbb{I}[h_V = v] \log(p(M_V = \perp \mid \boldsymbol{z}_V, \theta_V)). \end{aligned} \tag{34}$$

A.2 Regularized Optimization of DCBM

As discussed in Section 3.1, we regularize the model to avoid trivially deferring whenever the human is correct. In this way, we can account for the cost of deferring and relegating it to the most significative cases. We formalize this intuition by requiring the log-probability of deferring when the human is correct to be smaller then zero. Formally, we define the following constraint over all variables of the deferring system

$$\forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V = v] \log P(M_V = \perp \mid \boldsymbol{x}, \theta_V)\right] < 0 \tag{35}$$

$$\iff \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[-1 \cdot \mathbb{I}[h_V = v] \log P(M_V = \perp \mid \boldsymbol{x}, \theta_V)\right] > 0 \tag{36}$$

$$\iff \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V = v] \Psi(q(\boldsymbol{z}_V; \theta_V), \perp)\right] > 0. \tag{37}$$

In practice, we treat the constraint as a regularization term controlled by an hyperparameter $\lambda \in \mathbb{R}$. In particular, let

$$g_V(\boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y}) = \mathbb{I}[h_V = v] \Psi(q(\boldsymbol{z}_V; \theta_V), \perp), \tag{38}$$

be the value of the constraint on the variable $V \in \boldsymbol{V}$. We treat the constrained optimization problem as the following regularized unconstrained problem.

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\ell(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y})\right] - \lambda \sum_{V \in \boldsymbol{V}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[g_V(\boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y})\right] \tag{39}$$

$$= \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\ell(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y})\right] + \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[-\lambda \sum_{V \in \boldsymbol{V}} g_V(\boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y})\right] \tag{40}$$

$$= \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\ell(\boldsymbol{\theta} \mid \boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y}) - \lambda \sum_{V \in \boldsymbol{V}} g_V(\boldsymbol{c}, \boldsymbol{h}, \boldsymbol{x}, \boldsymbol{y})\right] \tag{41}$$

$$= \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + \mathbb{I}[h_V = v] \Psi(q(\boldsymbol{z}_V; \theta_V), \perp) - \lambda \mathbb{I}[h_V = v] \Psi(q(\boldsymbol{z}_V; \theta_V), \perp)\right] \tag{42}$$

$$= \min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + (1 - \lambda) \mathbb{I}[h_V = v] \Psi(q(\boldsymbol{z}_V; \theta_V), \perp)\right]. \tag{43}$$

Further, we show that the formulation from Liu et al. [2024] arises when explicitly constraining the model to avoid deferring whenever the human is incorrect in the training distribution as in $\mathbb{E}\left[\mathbb{I}[h_V \neq V] P(M_V \neq \perp \boldsymbol{x}, \theta_V)\right] > 0$. By expressing the constraint in terms of log-probabilities, we get the following result.

$$\forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V \neq v] \log P(M_V \neq \perp \mid \boldsymbol{x}, \theta_V)\right] > -\epsilon, \tag{44}$$

$$\iff \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V \neq v] \log \sum_{k \in [n_V]} P(M_V = k \mid \boldsymbol{x}, \theta_V)\right] > -\epsilon, \tag{45}$$

$$\impliedby \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V \neq v] \sum_{k \in [n_V]} \log P(M_V = k \mid \boldsymbol{x}, \theta_V)\right] > -\epsilon, \tag{46}$$

$$\iff \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[-1 \cdot \mathbb{I}[h_V \neq v] \sum_{k \in [n_V]} \log P(M_V = k \mid \boldsymbol{x}, \theta_V)\right] < \epsilon, \tag{47}$$

$$\iff \forall V \in \boldsymbol{V}. \quad \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \left[\mathbb{I}[h_V \neq v] \sum_{k \in [n_V]} \Psi(q(\boldsymbol{z}_V; \theta_V), v)\right] < \epsilon, \tag{48}$$

Loss Name	Loss Function
CE [Mozannar et al., 2023]	$\psi\left(q(z), k\right) = -\log\left(\frac{\exp(q(z)_k)}{\sum_{k' \in [K+1]} \exp(q(z)_{k'})}\right)$
OVA [Verma et al., 2023]	$\psi\left(q(z), k\right) = \begin{cases} \log\left(1+\exp\left(-q(z)_k\right)\right) - \log\left(1+\exp\left(+q(z)_k\right)\right) & \text{if } k = \perp \\ \log\left(1+\exp\left(-q(z)_k\right)\right) + \sum_{k' \in [K+1]/\{k\}} \log\left(1+\exp\left(+q(z)_{k'}\right)\right) & \text{otherwise} \end{cases}$
ASM [Cao et al., 2023]	$\psi\left(q(z), k\right) = \begin{cases} -\log\left(\frac{\exp(q(z)_k)}{\sum_{k' \in [K]} \exp(q(z)_{k'}) - \max_{k' \in [K]} \exp(q(z)_{k'})}\right) & \text{if } k = \perp \\ -\log\left(\frac{\exp(q(z)_k)}{\sum_{k' \in [K]} (\exp(q(z)_{k'})}\right) - \log\left(\frac{\sum_{k' \in [K]} \exp(q(z)_{k'}) - \max_{k' \in [K]} \exp(q(z)_{k'})}{\sum_{k' \in [K+1]} \exp(q(z)_{k'}) - \max_{k' \in [K]} \exp(q(z)_{k'})}\right) & \text{otherwise} \end{cases}$

Table 1: Multiclass losses from Liu et al. [2024].

for a positive threshold $\epsilon > 0$. Consequently, when introducing this constraints with the same penalty λ in the optimization problem, we obtain the following formulation

$$\min_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{c},\boldsymbol{h},\boldsymbol{x},\boldsymbol{y}} \Bigg[\sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + (1-\lambda)\mathbb{I}[h_V = v]\Psi(q(\boldsymbol{z}_V; \theta_V), \perp) + \lambda \mathbb{I}[h_V \neq v] \sum_{k \in [n_V]} \Psi(q(\boldsymbol{z}_V; \theta_V), k) \Bigg] \tag{49}$$

A.3 Lemma 3.2 — Sum of Consistent Losses

Before proving Lemma 3.2, we prove the following result that is a fundamental property arising from the definition of the argmin function and the independence of variables.

Lemma A.1. *Given $f, g : A \subseteq \mathbb{R}^n \to \mathbb{R}$, we have*

$$\arg\min_{(x,y) \in A^2} \left(f(x) + g(y)\right) = \{(\bar{x}, \bar{y}) \in A^2 : \ \bar{x} \in \arg\min_{x \in A} f(x), \bar{y} \in \arg\min_{y \in A} g(y)\}$$

Proof. For the sake of simplicity we use the following shortcut:

$$L = \arg\min_{(x,y) \in A^2} \left(f(x) + g(y)\right) \quad \text{and} \quad R = \{(\bar{x}, \bar{y}) \in A^2 : \ \bar{x} \in \arg\min_{x \in A} f(x) \wedge \bar{y} \in \arg\min_{y \in A} g(y)\}$$

First, we notice that in case any between f or g has no minimum in A, then the claim is trivially proved as $L = R = \emptyset$. Indeed, let us assume, e.g., that f has no minimum in A, then clearly $R = \emptyset$. Moreover, also $L = \emptyset$. Indeed, if we assume by contradiction that $L \neq \emptyset$, then it exists $(\bar{x}, \bar{y}) \in L$, i.e. $(\bar{x}, \bar{y}) \in A^2$ with $f(\bar{x}) + g(\bar{y}) \leq f(x) + g(y)$ for every $(x, y) \in A^2$. By taking $y = \bar{y}$ and canceling $g(\bar{y})$ on both sides we get that $f(\bar{x}) \leq f(x)$ for every $x \in A$. Therefore f has at least a minimum ($\bar{x}$) in A, which is a contradiction, so it must be $L = \emptyset$, as well.

So lets consider the case of both $L \neq \emptyset$ and $R \neq \emptyset$. We show the double inclusion.

1. If $(\bar{x}, \bar{y}) \in L$ then $f(\bar{x}) + g(\bar{y}) \leq f(x) + g(y)$ for every $(x, y) \in A^2$. From this inequality, by taking $x = \bar{x}$ and canceling $f(\bar{x})$, we get $\bar{y} \in \arg\min_{y \in A} g(y)$. Identically, by taking $y = \bar{y}$, we get $\bar{x} \in \arg\min_{x \in A} f(x)$. Therefore $(\bar{x}, \bar{y}) \in R$.
2. If $(\bar{x}, \bar{y}) \in R$ then $\bar{x} \in \arg\min_{x \in A} f(x)$ and $\bar{y} \in \arg\min_{y \in A} g(y)$. Namely, $f(\bar{x}) \leq f(x)$ for every $x \in A$ and $g(\bar{y}) \leq g(y)$ for every $y \in A$. By summing on both sides, we get $f(\bar{x}) + g(\bar{y}) \leq f(x) + g(y)$ for every $(x, y) \in A^2$, and so $(\bar{x}, \bar{y}) \in L$

□

Lemma 3.2 Let $\ell'_1, \ell_1, \cdots, \ell'_m, \ell_m$ be (possibly distinct) loss functions. Assume that, for every $i \in \{1, \ldots, m\}$, $\ell'_i, \ell_i : \mathbb{R}^{n_i} \to \mathbb{R}$, being ℓ'_i a consistent surrogate of ℓ_i. Then $\ell' : \mathbb{R}^n \to \mathbb{R}$, with $n = n_1 + \ldots + n_m$ and $\ell'(\theta_1, \ldots, \theta_m) = \sum_{i=1}^m \ell'_i(\theta_i)$ is a consistent surrogate of $\ell : \mathbb{R}^n \to \mathbb{R}$, with $\ell(\theta_1, \ldots, \theta_m) = \sum_{i=1}^m \ell(\theta_i)$.

Proof. The proof is a direct consequence of Lemma A.1. For simplicity, we show the complete proof for $m = 2$. To be precise, from the statement we report explicitly that, $\ell'_1, \ell_1 : \mathbb{R}^{n_1} \to \mathbb{R}$, $\ell'_2, \ell_2 : \mathbb{R}^{n_2} \to \mathbb{R}$ and $\ell', \ell : \mathbb{R}^n \to \mathbb{R}$ with $n = n_1 + n_2$, $\ell' = \ell'_1 + \ell'_2$ and $\ell = \ell_1 + \ell_2$. We have to prove that ℓ' is a consistent surrogate of ℓ, namely that $\arg\min_{\theta \in \mathbb{R}^n} \ell'(\theta) \subseteq \arg\min_{\theta \in \mathbb{R}^n} \ell(\theta)$.

Let $\theta^* = (\theta^*_1, \theta^*_2) \in \mathbb{R}^{n_1+n_2}$ be a minimum of ℓ' (the claim would be trivial in case ℓ' has no minima). Then according to Lemma A.1, we have:

$$\begin{aligned} \theta^* \in \arg\min_{\theta \in \mathbb{R}^n} \ell'(\theta) = \arg\min_{(\theta_1,\theta_2) \in \mathbb{R}^{n_1+n_2}} (\ell'_1(\theta_1) + \ell'_2(\theta_2)) = \\ \{(\bar{\theta}_1, \bar{\theta}_2) \in \mathbb{R}^{n_1+n_2} : \bar{\theta}_1 \in \arg\min_{\theta_1 \in \mathbb{R}^{n_1}} \ell'_1(\theta_1) \wedge \bar{\theta}_2 \in \arg\min_{\theta_2 \in \mathbb{R}^{n_2}} \ell'_2(\theta_2)\} \end{aligned} \tag{50}$$

Therefore $\theta^*_1 \in \arg\min_{\theta_1 \in \mathbb{R}^{n_1}} \ell'_1(\theta_1)$ and $\theta^*_2 \in \arg\min_{\theta_2 \in \mathbb{R}^{n_2}} \ell'_2(\theta_2)$. Since by hypothesis ℓ'_1, ℓ'_2 are consistent surrogates of ℓ_1, ℓ_2, respectively, it follows that: $\theta^*_1 \in \arg\min_{\theta_1 \in \mathbb{R}^{n_1}} \ell_1(\theta_1)$ and $\theta^*_2 \in \arg\min_{\theta_2 \in \mathbb{R}^{n_2}} \ell_2(\theta_2)$.

Finally, the proof concludes by using again Lemma A.1:

$$\begin{aligned} \theta^* \in \{(\bar{\theta}_1, \bar{\theta}_2) \in \mathbb{R}^{n_1+n_2} : \bar{\theta}_1 \in \arg\min_{\theta_1 \in \mathbb{R}^{n_1}} \ell_1(\theta_1) \wedge \bar{\theta}_2 \in \arg\min_{\theta_2 \in \mathbb{R}^{n_2}} \ell_2(\theta_2)\} = \\ \arg\min_{(\theta_1,\theta_2) \in \mathbb{R}^{n_1+n_2}} (\ell_1(\theta_1) + \ell_2(\theta_2)) = \arg\min_{\theta \in \mathbb{R}^n} \ell(\theta) \end{aligned} \tag{51}$$

□

A.4 Theorem 1 — Sum of Consistent Losses

Proof. By the previous Lemma 3.2, the sum of consistent losses is consistent to the sum of the target loss functions. It is thus immediate how this applies our optimization problem both for the unconstrained (Equation 3) and the penalized (Equation 4) losses. Formally,

$$\sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + (1 - \lambda)\mathbb{I}[h_V = v]\Psi(q(\boldsymbol{z}_V; \theta_V), \perp) \tag{52}$$

is the sum of losses consistent of the zero-one loss which we reported in Equation 1 whenever $\lambda = 1$. In fact, it corresponds to an equivalent formulation in Theorem 1 from Mozannar and Sontag [2020]. Similarly, for any other $\lambda \in [0, 1]$, the penalized version coincides in the single-variable case to the provably consistent formulation from Equation 4 in Liu et al. [2024]. Therefore, the sum over different variables is consistent to the sum of the zero-one loss. □

B Experimental Details

Data Split. For the `completeness` synthetic dataset, we sample $1,000$ instances with an 80%-20% train-test split ratio. For `cifar10h`, we randomly split the dataset into training, validation and test according to a $70\%, 10\%, 20\%$ ratio. For `CUB`, we keep the original split.

Architecture of Concept and Task Predictors For each concept predictor q_C, we employ a three-layer MLP with a leaky-relu activation function. The black-box baselines and the CBM models, including our DCBM, adopt the same architecture and the same common frozen representation. Then, the task and concept classifiers are trained independently for the black box, the standard CBM, our DCBM, and its ablations (DCBM-NoTask and DCBM-NoConcepts) For the `completeness` dataset, each concept encoder model takes as input the raw data. For the image datasets `cifar10-h` and `cub`, concept predictors take instead as an input the pre-trained embedding discussed in Section 3.4. For `CUB`, we obtain such an embedding by training a ResNet34 [He et al., 2016] for 100 epochs to solve the final task using a cross-entropy loss function. The representations obtained by the pre-trained model are then frozen and used as the input for each concept encoder. For `cifar10-h` we consider the pre-trained WideResNet [Zagoruyko and Komodakis, 2016] provided by Palomba et al. [2025], who trained a WideResNet architecture on the original `cifar10` [Krizhevsky et al., 2009] training set for 200 epochs. We use the obtained representations to train all the concept encoders. All the final task classifiers consist of another three-layer MLP taking as input the concept values.

Table 2: Mapping for uncertainty of concepts by Koh et al. [2020].

c	1	1	1	1	0	0	0	0
u	1	2	3	4	1	2	3	4
$p_H(\hat{c} \mid c, u)$	0.00	0.50	0.75	1.00	0.00	0.50	0.25	0.00

Uncertain Concepts. To produce human expert labels in the `CUB` dataset, we employ the following strategy. Let c be the ground-truth label of a concept for a given sample and u be the corresponding label of uncertainty, as provided by Koh et al. [2020]. Uncertainty labels have the following semantics: not visible ($u = 1$), guessing ($u = 2$), probably ($u = 3$), and definitely ($u = 4$). Koh et al. [2020] translate the uncertainty labels in the following probabilities, which we use to sample the value $\hat{c}$ of the concept provided by a human practitioner.

Training Procedure. We train every combination of models and defer costs λ for 100 epochs. For `completeness`, we use `Adam` [Kingma and Ba, 2015] with a learning rate equal to $.001$ and no scheduler. For both `cifar10-h` and `CUB`, we use `AdamW` [Loshchilov and Hutter, 2019] as an optimizer, setting the initial learning rate to $.001$. We decrease the learning rate every 25 epochs by $.5$. Additionally, for `CUB`, following Zarlenga et al. [2022] guidelines, we consider a weighted version of the loss on concepts to take into account their imbalance. To limit the computational burden, for both `cifar10-h` and `CUB`, we perform early stopping after 10 epochs if there is no improvement for the loss on the validation set.

Evaluation. All the results are averaged over five and three runs on the synthetic and the other datasets, respectively, with fixed datasets' splits.

Hardware and Computational Time We train our baselines on a 224 cores machine with Intel(R) Xeon(R) Platinum 8480+ CPU and eight NVIDIA A100-SXM4-80GB, OS Ubuntu 22.04.4 LTS. Notably, the cost of training a DCBM against a CBM is negligible, as they share essentially the same architecture, apart for an additional feature on each concept or task classifier. In detail, on our hardware, for a training epoch it takes ≈ 9 seconds on `CUB` for a DCBM and ≈ 7.5 for a standard CBM. For `cifar10-h`, given the smaller number of concepts, the difference is even more negligible, with both taking approximately one second per epoch. We report epoch time as, due to early stopping strategies, the overall training time might vary across runs.

C Additional Explanations

In Figure 5, we report additional samples from `CUB` that we analyze as in the experimental results for research question Q4 in Section 4.2. While in the first two rows the effect of deferring on concepts is clearly beneficial, the other two examples require further discussion.

In both cases, we can see that the human would still mispredict a few concepts. Interestingly, such concepts are not visible in the image (e.g., the upper tail in the second-to-last example and the belly colour in the last example). Therefore, the perfect interventions are just an ideal scenario, as in practice, these concepts are not directly predictable from the image. Still, `DCBM` can "correctly defer" on the final task as concept interventions would not suffice to disambiguate the correct label.

D Additional Comparisons

D.1 Defer on Task

Deferring on the final task is a useful strategy to mitigate risk in incomplete scenarios where concepts alone are insufficient to determine the task label. For instance, in the `cifar-10h` dataset, concepts do not allow distinguishing cats from deer, since they have the same concept-level representation. In these cases, incomplete concept combinations should trigger the defer option on the task, highlighting

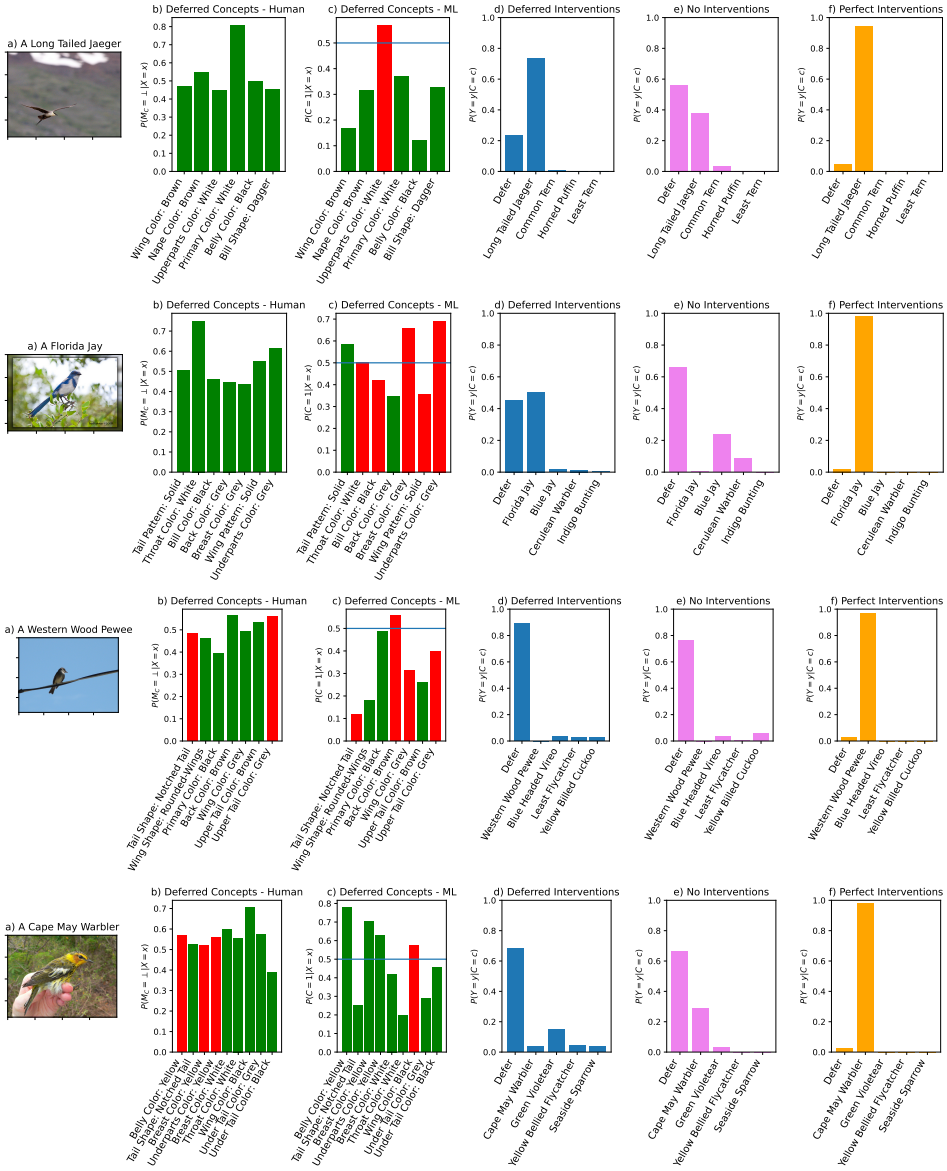


Figure 5: Interpretation of a DCBM with defer cost $\lambda = 0.1$ on an input sample. From left to right: (a.) examples of an image from the CUB dataset; (b.) the concepts that the model has deferred with the estimated probability, green bars stand for when the human correctly predicts the concept, red otherwise; (c.) the estimated probability of each deferred concept being true according to the machine learning model, green bars stand for when the ML would have correctly predicted the concept, red otherwise; (d.) the estimated probability of top-5 final task labels after deferring the concepts to the human (standard DCBM behavior); (e.) the estimated probability of top-5 final task labels without deferring the concepts to the human; (f.) the estimated probability of top-5 final task labels from the ground-truth concepts;

Table 3: Coverage of all concepts for increasing cost λ on `cifar10-h`. The DCBM correctly defers cats and deer when defer is not too costly, improving final performance. We highlight in bold the classes that are not possible to distinguish based on concept representations, i.e., `cat` and `deer`

λ	plane	auto	bird	**cat**	**deer**	dog	frog	horse	ship	truck
0.00	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$\mathbf{.000 \pm .000}$	$\mathbf{.000 \pm .000}$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$
0.01	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$\mathbf{.000 \pm .000}$	$\mathbf{.000 \pm .000}$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$
0.05	$.988 \pm .003$	$.998 \pm .003$	$.982 \pm .003$	$\mathbf{.032 \pm .014}$	$\mathbf{.009 \pm .003}$	$.999 \pm .003$	$.984 \pm .007$	$.979 \pm .007$	$.987 \pm .003$	$.989 \pm .003$
0.10	$.978 \pm .010$	$.998 \pm .003$	$.982 \pm .006$	$\mathbf{.047 \pm .009}$	$\mathbf{.015 \pm .003}$	$.981 \pm .011$	$.978 \pm .005$	$.977 \pm .007$	$.977 \pm .018$	$.984 \pm .008$
0.20	$.970 \pm .023$	$.998 \pm .003$	$.980 \pm .022$	$\mathbf{.126 \pm .014}$	$\mathbf{.037 \pm .017}$	$.957 \pm .025$	$.984 \pm .003$	$.982 \pm .010$	$.986 \pm .005$	$.970 \pm .000$
0.30	$.983 \pm .003$	$.996 \pm .003$	$.983 \pm .012$	$\mathbf{.181 \pm .013}$	$\mathbf{.064 \pm .003}$	$.961 \pm .007$	$.989 \pm .010$	$.981 \pm .005$	$.990 \pm .000$	$.980 \pm .009$
0.40	$.987 \pm .010$	$1.00 \pm .000$	$.987 \pm .006$	$\mathbf{.739 \pm .394}$	$\mathbf{.676 \pm .478}$	$.978 \pm .020$	$.986 \pm .016$	$.992 \pm .010$	$.998 \pm .003$	$.995 \pm .005$
0.50	$.998 \pm .003$	$1.00 \pm .000$	$.995 \pm .000$	$\mathbf{.975 \pm .011}$	$\mathbf{.979 \pm .018}$	$.997 \pm .005$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$.997 \pm .003$

Table 4: $CovConc$ for the first 5 concepts, where the human is always correct. `DCBM-NT` always defer when possible ($CovConc$ is close to zero for small λ), while `UCP` fails and defer less than optimal.

	$CovConc - 0$		$CovConc - 1$		$CovConc - 2$		$CovConc - 3$		$CovConc - 4$	
λ	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT
0.00	$.615 \pm .117$	$.000 \pm .000$	$.676 \pm .091$	$.006 \pm .013$	$.584 \pm .091$	$.000 \pm .000$	$.708 \pm .141$	$.002 \pm .004$	$.316 \pm .083$	$.000 \pm .000$
0.01	$.619 \pm .111$	$.017 \pm .019$	$.680 \pm .088$	$.030 \pm .045$	$.589 \pm .088$	$.017 \pm .019$	$.713 \pm .137$	$.003 \pm .007$	$.325 \pm .084$	$.000 \pm .000$
0.05	$.669 \pm .097$	$.132 \pm .084$	$.724 \pm .091$	$.177 \pm .110$	$.631 \pm .077$	$.143 \pm .079$	$.741 \pm .139$	$.110 \pm .055$	$.394 \pm .090$	$.000 \pm .000$
0.10	$.777 \pm .039$	$.382 \pm .162$	$.810 \pm .087$	$.529 \pm .073$	$.677 \pm .082$	$.262 \pm .131$	$.792 \pm .111$	$.180 \pm .085$	$.513 \pm .149$	$.071 \pm .081$
0.25	$.932 \pm .06$	$.890 \pm .019$	$.94 \pm .053$	$.943 \pm .018$	$.854 \pm .054$	$.714 \pm .063$	$.958 \pm .034$	$.838 \pm .053$	$.907 \pm .093$	$.459 \pm .165$
0.50	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$

instances that cannot be classified by looking at concepts only. We show this experimentally on `cifar-10h`, where we can see that deferring on the final task improves the final accuracy (Figure 3c). Instead, the accuracy of the ablated `DCBM-NT` (i.e., a DCBM with no possibility to defer on the final task) plateaus at around 90%. This is due to `DCBM-NT` randomly guessing between deer and cats, while correctly classifying other classes. Furthermore, Table 3 reports the coverage of all concepts for increasing cost λ. Results show that, whenever defer is not too costly, `DCBM` correctly defers only instances of cats and deer to a human. Coherently with our formulation, when the cost increases, the model instead prefers to take a guess instead of deferring.

D.2 Uncertainty on Concept Predictions (UCP) vs DCBMs

Intervention strategies, such as those proposed by Shin et al. [2023], allocate a number of admissible interventions and then choose for each instance (or for a batch of instances) on which concepts to intervene. Typically, such intervention strategies only consider the uncertainty of the model, which might disregard the fact that a human would not be better than the ML predictor in classifying a particular instance. Using learning to defer methodologies, we instead equip CBMs with the capability to (i.) autonomously ask for human intervention and (ii.) acknowledge the capabilities of the expert, i.e., we consider fallible human beings with variable performance.

We validate this intuition through two extra experiments, i.e., one in a fully controlled setting and one over `CUB` with uncertain humans. For both experiments, we compare `DCBM-NT` - i.e., a DCBM without the option to defer on the final task - and the Uncertainty of Concept Predictions strategy (`UCP`) [Shin et al., 2023], which determines when to abstain based on the uncertainty on the concepts.

Synthetic Example We consider a slight modification of the `completeness` data we use for our ablation study (see Appendix E): we define a scenario where the human is always correct, on the first 5 concepts (out of 10) and always wrong on the remaining 5 concepts, i.e., the concept predictions for these last 5 concepts always differ from the ground truth concepts.

We report in Tables 4 and 5 results for the coverage over the 10 concepts for both `DCBM-NT` and `UCP` applied on top of an independent CBM. Recall that on the first 5 concepts the human is always correct, hence, if the cost allows it, `DCBM-NT` correctly learns to defer ($CovConc$ is 0 at $\lambda = 0$), as shown in Table 4.

Conversely, on the last 5 concepts, where the human always makes mistakes (Table 5) we can see that the coverage for `DCBM-NT` is always one, i.e., the model has learned that interventions there would be harmful. On the other hand, `UCP` coverage is below one, meaning the intervention strategy would require intervening on concepts where the human is wrong.

Table 5: $CovConc$ for the last five concepts, where the human expert is always wrong. DCBM-NT correctly never defers ($CovConc$ is one always), while UCP fails and defer more than ideal.

	$CovConc-5$		$CovConc-6$		$CovConc-7$		$CovConc-8$		$CovConc-9$	
λ	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT	UCP	DCBM-NT
0.00	$.346 \pm .127$	$1.00 \pm .000$	$.279 \pm .203$	$1.00 \pm .000$	$.763 \pm .060$	$1.00 \pm .000$	$.322 \pm .056$	$1.00 \pm .000$	$.399 \pm .174$	$1.00 \pm .000$
0.01	$.360 \pm .134$	$1.00 \pm .000$	$.281 \pm .205$	$1.00 \pm .000$	$.768 \pm .061$	$1.00 \pm .000$	$.328 \pm .054$	$1.00 \pm .000$	$.404 \pm .173$	$1.00 \pm .000$
0.05	$.416 \pm .154$	$1.00 \pm .000$	$.325 \pm .205$	$1.00 \pm .000$	$.805 \pm .050$	$1.00 \pm .000$	$.388 \pm .045$	$1.00 \pm .000$	$.469 \pm .146$	$1.00 \pm .000$
0.10	$.506 \pm .174$	$1.00 \pm .000$	$.405 \pm .213$	$1.00 \pm .000$	$.881 \pm .033$	$1.00 \pm .000$	$.482 \pm .053$	$1.00 \pm .000$	$.581 \pm .127$	$1.00 \pm .000$
0.25	$.774 \pm .152$	$1.00 \pm .000$	$.789 \pm .070$	$1.00 \pm .000$	$.973 \pm .028$	$1.00 \pm .000$	$.848 \pm .086$	$1.00 \pm .000$	$.869 \pm .108$	$1.00 \pm .000$
0.50	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$	$1.00 \pm .000$

Table 6: Comparison between DCBM-NT and UCP over CUB with uncertain humans. Results show that DCBM-NT is able to require intervention only when needed, while UCP over relies on humans, even if these can make mistakes.

λ	$CovConc$	UCP	DCBM-NT
0.00	$.185 \pm .002$	$.665 \pm .015$	$\mathbf{.800 \pm .004}$
0.01	$.188 \pm .002$	$.666 \pm .015$	$\mathbf{.803 \pm .000}$
0.02	$.271 \pm .029$	$.685 \pm .010$	$\mathbf{.796 \pm .001}$
0.0225	$.424 \pm .049$	$.715 \pm .024$	$\mathbf{.782 \pm .003}$
0.025	$.601 \pm .033$	$.746 \pm .012$	$\mathbf{.756 \pm .002}$
0.0275	$.700 \pm .057$	$\mathbf{.761 \pm .015}$	$.743 \pm .007$
0.03	$.808 \pm .016$	$\mathbf{.771 \pm .002}$	$.738 \pm .004$
0.04	$.916 \pm .005$	$\mathbf{.751 \pm .001}$	$.724 \pm .002$
0.05	$.931 \pm .001$	$\mathbf{.745 \pm .003}$	$.723 \pm .004$
0.10	$.963 \pm .002$	$\mathbf{.719 \pm .003}$	$.704 \pm .004$
0.20	$.990 \pm .002$	$\mathbf{.690 \pm .001}$	$.685 \pm .001$
0.30	$.998 \pm .001$	$\mathbf{.682 \pm .002}$	$.677 \pm .002$
0.50	$1.000 \pm .000$	$\mathbf{.679 \pm .002}$	$.676 \pm .003$

CUB with uncertain humans Table 6 reports the comparison over CUB with uncertain humans.

The results show that when we allow for a large number of interventions ($\lambda \leq .025$), UCP underperforms because it asks for interventions based solely on model uncertainty, without accounting for whether the human is likely to be more accurate. As a result, it defers to the human on instances where the ML model predictions would have been a better option.

On the other hand, when the budget of interventions is limited, DCBMs have a more conservative approach and tend to prefer the use of the ML model, leading to slightly worse outcomes.

E Additional Results

The completeness dataset is a synthetic dataset that allows full control of the data-generation process [Laguna et al., 2024]. We add labels from human experts with different competencies by selecting the concepts' or task's correct labels with different probabilities. In particular, we denote as oracle, human-80% and human-60%, a human that correctly predicts their labels with an accuracy of 100%, 80% and 60%, respectively. For the oracle scenario, we plot the results in Figure 6.

We provide additional results on this dataset to investigate the following comparisons:

- Shared parameters among the concept encoders,
- Different human expert accuracy on both the concepts and task,
- Joint vs Independent training,
- Different learning-to-defer losses.

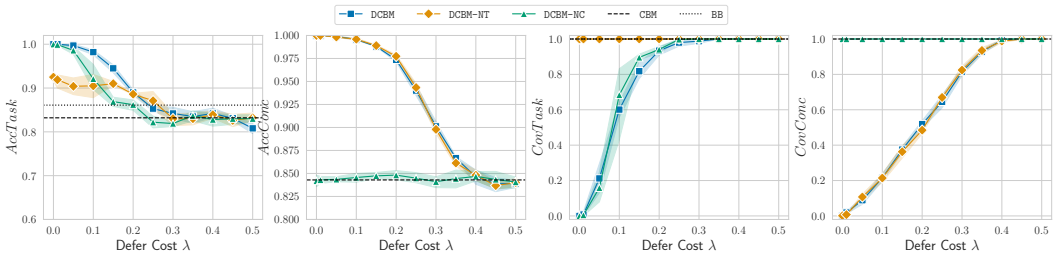


Figure 6: Results on `completeness` when human experts have perfect concept and task accuracy (i.e., they are oracles). We report each metric's average and standard deviations as we increase the defer cost λ. The black box and the CBM baselines are constant as they are independent of the defer cost.

Label Smoothing. Liu et al. [2024] studies the problem of label smoothing on learning to defer losses a proposes a slightly different formulation of Equation 4, which, once adapted to our notation, we report as follows:

$$\begin{aligned}\sum_{V \in \boldsymbol{V}} \Psi(q(\boldsymbol{z}_V; \theta_V), v) + (1-\lambda) \cdot \mathbb{I}\left[y_V = h_V\right] \Psi(q(\boldsymbol{z}_V; \theta_V), \bot) \\ + \lambda \cdot \mathbb{I}\left[y_V \neq h_V\right] \arg\min_{k \in [K]} \Psi(q(\boldsymbol{z}_V; \theta_V), k),\end{aligned} \tag{53}$$

In all the coming experimental results, we use the suffix `-LS` to refer to the results using Equation 53, while we use the suffix `-NLS` for the formulation in Equation 4. As it is shown in the upcoming additional results, we do not observe noteworthy differences in the performance of the two loss functions.

Joint Learning. While independent training is required to guarantee the consistency of the learning-to-defer loss function, we implement joint learning to compare empirically. We implement the joint learning strategy by considering the following soft-labelled concept predictor:

$$\tilde{g}(\boldsymbol{x}) = g_1(\boldsymbol{x})(1 - g_\bot(\boldsymbol{x})) + h_c(g_\bot(\boldsymbol{x})), \tag{54}$$

where we produce the output as the weighted sum of the human-provided concept ψ and the machine learning model concept. The weight corresponds to the probability of deferring or not the instance.

We study the different negative log-likelihood terms (Table 1) that can be employed within the learning-to-defer (Equation 4) loss function for both independent (Tables 23, 27 and 29) and joint learning (Tables 24, 28 and 30), when dealing with oracle human experts on both concepts and tasks. Further, for the ASM loss function, we also study how the model behaves when we do not freeze the parameters of the encoder, also for independent (Table 25) and joint training (Table 26). Finally, we study multiple combinations of human expertise on the concepts and the tasks, whose reference we summarize in the following table:

		Human Task Expert		
		60	80	oracle
Human Concept Expert	60	Tables 7 and 8	Tables 9 and 10	Tables 11 and 12
	80	Tables 13 and 14	Tables 15 and 16	Tables 17 and 18
	oracle	Tables 19 and 20	Tables 21 and 22	Tables 23 and 24

Table 7: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human60 task expert and human60 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	**.828 ± .024**	.815 ± .008	.826 ± .019	.827 ± .021	.828 ± .010	.816 ± .014
	0.01	.819 ± .017	.825 ± .015	.825 ± .009	.821 ± .018	**.831 ± .017**	.819 ± .021
	0.05	.813 ± .015	**.836 ± .013**	.833 ± .008	.822 ± .018	.833 ± .014	.819 ± .014
	0.10	.824 ± .022	.814 ± .014	.832 ± .014	.822 ± .010	**.834 ± .014**	.829 ± .023
	0.15	.828 ± .008	.828 ± .016	.824 ± .015	.824 ± .018	.818 ± .013	**.829 ± .008**
	0.20	.819 ± .016	.825 ± .014	**.839 ± .020**	.834 ± .011	.834 ± .016	**.839 ± .014**
	0.25	.823 ± .006	**.836 ± .011**	.819 ± .016	.823 ± .016	.817 ± .019	.815 ± .023
	0.30	.828 ± .012	**.831 ± .016**	.828 ± .016	.822 ± .014	.821 ± .010	.819 ± .010
	0.35	.814 ± .014	.822 ± .014	.823 ± .021	**.839 ± .010**	.831 ± .019	.829 ± .013
	0.40	**.834 ± .023**	.824 ± .019	.825 ± .016	.830 ± .015	**.834 ± .016**	.831 ± .012
	0.45	.826 ± .010	**.835 ± .014**	.833 ± .024	.823 ± .008	.827 ± .016	.824 ± .018
	0.50	.808 ± .015	.818 ± .012	.821 ± .015	.810 ± .009	**.828 ± .006**	.814 ± .024
$AccConc$	0.00	.830 ± .009	**.845 ± .008**	.829 ± .007	.826 ± .007	.842 ± .008	.833 ± .010
	0.01	.827 ± .012	**.843 ± .010**	.831 ± .007	.831 ± .009	.843 ± .005	.827 ± .007
	0.05	.824 ± .006	**.845 ± .007**	.824 ± .012	.828 ± .011	.844 ± .003	.829 ± .008
	0.10	.833 ± .005	.843 ± .010	.832 ± .008	.822 ± .008	**.846 ± .006**	.834 ± .008
	0.15	.828 ± .008	.847 ± .008	.826 ± .009	.834 ± .008	**.847 ± .004**	.834 ± .009
	0.20	.823 ± .011	.844 ± .006	.830 ± .009	.826 ± .010	**.848 ± .007**	.831 ± .006
	0.25	.822 ± .008	.845 ± .003	.826 ± .008	.826 ± .007	**.845 ± .007**	.827 ± .015
	0.30	.827 ± .007	**.849 ± .008**	.826 ± .006	.821 ± .009	.841 ± .007	.824 ± .006
	0.35	.821 ± .011	**.845 ± .004**	.817 ± .010	.830 ± .006	.844 ± .012	.824 ± .006
	0.40	.821 ± .009	.842 ± .016	.824 ± .008	.825 ± .011	**.847 ± .007**	.815 ± .019
	0.45	.823 ± .010	**.845 ± .012**	.818 ± .017	.823 ± .006	.844 ± .012	.821 ± .013
	0.50	.812 ± .008	**.851 ± .007**	.816 ± .006	.814 ± .003	.841 ± .008	.817 ± .008
$CovTask$	0.00	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.01	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.05	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.10	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.15	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.20	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.25	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.30	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.35	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.40	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.45	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
$CovConc$	0.00	.973 ± .018	**1.000 ± 0.000**	.971 ± .017	.975 ± .013	**1.000 ± 0.000**	.979 ± .020
	0.01	.983 ± .016	**1.000 ± 0.000**	.971 ± .033	.976 ± .012	**1.000 ± 0.000**	.984 ± .003
	0.05	.991 ± .017	**1.000 ± 0.000**	.997 ± .004	.999 ± .002	**1.000 ± 0.000**	.999 ± .002
	0.10	**1.000 ± 0.000**	**1.000 ± 0.000**	.999 ± .002	1.000 ± .001	**1.000 ± 0.000**	1.000 ± 0.000
	0.15	**1.000 ± 0.000**	**1.000 ± 0.000**	1.000 ± 0.000	**1.000 ± 0.000**	**1.000 ± 0.000**	1.000 ± 0.000
	0.20	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.25	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.30	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.35	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.40	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.45	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**

Table 8: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human60 task expert and human60 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.822 \pm .004$	$.831 \pm .009$	$\mathbf{.840 \pm .010}$	$.827 \pm .016$	$.826 \pm .031$	$.827 \pm .019$
	0.01	$.816 \pm .011$	$.818 \pm .013$	$.815 \pm .008$	$.816 \pm .014$	$\mathbf{.821 \pm .012}$	$.820 \pm .006$
	0.05	$\mathbf{.829 \pm .018}$	$.820 \pm .018$	$.823 \pm .012$	$.821 \pm .014$	$.828 \pm .014$	$.827 \pm .017$
	0.10	$\mathbf{.830 \pm .012}$	$.829 \pm .007$	$.814 \pm .011$	$.814 \pm .013$	$.828 \pm .007$	$.824 \pm .022$
	0.15	$.813 \pm .012$	$\mathbf{.837 \pm .010}$	$.829 \pm .007$	$.829 \pm .012$	$.835 \pm .013$	$.828 \pm .010$
	0.20	$.824 \pm .008$	$\mathbf{.834 \pm .014}$	$.822 \pm .017$	$.827 \pm .010$	$.824 \pm .021$	$.816 \pm .015$
	0.25	$.817 \pm .006$	$.829 \pm .012$	$.814 \pm .008$	$.823 \pm .018$	$\mathbf{.833 \pm .008}$	$.831 \pm .022$
	0.30	$.824 \pm .010$	$\mathbf{.833 \pm .008}$	$.827 \pm .014$	$\mathbf{.833 \pm .016}$	$.829 \pm .018$	$.822 \pm .008$
	0.35	$.813 \pm .019$	$\mathbf{.836 \pm .018}$	$.826 \pm .009$	$.820 \pm .004$	$.829 \pm .011$	$\mathbf{.836 \pm .008}$
	0.40	$.817 \pm .014$	$\mathbf{.831 \pm .005}$	$.820 \pm .013$	$.828 \pm .010$	$.827 \pm .008$	$.823 \pm .007$
	0.45	$.803 \pm .034$	$.815 \pm .013$	$.817 \pm .025$	$.809 \pm .025$	$.816 \pm .016$	$\mathbf{.825 \pm .021}$
	0.50	$.800 \pm .057$	$.821 \pm .011$	$.809 \pm .014$	$.815 \pm .007$	$.825 \pm .005$	$\mathbf{.832 \pm .008}$
$AccConc$	0.00	$.833 \pm .007$	$\mathbf{.841 \pm .010}$	$.817 \pm .004$	$.828 \pm .011$	$.832 \pm .005$	$.821 \pm .011$
	0.01	$.825 \pm .008$	$.837 \pm .008$	$.822 \pm .007$	$.821 \pm .005$	$\mathbf{.843 \pm .012}$	$.830 \pm .011$
	0.05	$.826 \pm .004$	$.836 \pm .009$	$.824 \pm .012$	$.821 \pm .015$	$\mathbf{.839 \pm .008}$	$.827 \pm .011$
	0.10	$.827 \pm .007$	$\mathbf{.839 \pm .004}$	$.826 \pm .005$	$.825 \pm .008$	$.831 \pm .006$	$.820 \pm .003$
	0.15	$.828 \pm .007$	$.826 \pm .007$	$\mathbf{.834 \pm .014}$	$.824 \pm .013$	$.831 \pm .011$	$.830 \pm .009$
	0.20	$.822 \pm .008$	$.839 \pm .007$	$.824 \pm .018$	$.821 \pm .010$	$\mathbf{.843 \pm .011}$	$.825 \pm .013$
	0.25	$.831 \pm .007$	$.841 \pm .005$	$.823 \pm .007$	$.824 \pm .012$	$\mathbf{.844 \pm .009}$	$.819 \pm .012$
	0.30	$.823 \pm .011$	$\mathbf{.845 \pm .010}$	$.822 \pm .009$	$.821 \pm .006$	$.834 \pm .011$	$.823 \pm .012$
	0.35	$.816 \pm .009$	$\mathbf{.839 \pm .006}$	$.808 \pm .007$	$.821 \pm .004$	$.835 \pm .004$	$.815 \pm .006$
	0.40	$.812 \pm .006$	$\mathbf{.841 \pm .012}$	$.808 \pm .011$	$.822 \pm .009$	$.839 \pm .005$	$.808 \pm .017$
	0.45	$.816 \pm .006$	$.837 \pm .010$	$.817 \pm .011$	$.825 \pm .010$	$\mathbf{.838 \pm .007}$	$.818 \pm .009$
	0.50	$.814 \pm .021$	$\mathbf{.836 \pm .013}$	$.804 \pm .011$	$.812 \pm .006$	$.832 \pm .008$	$.811 \pm .007$
$CovTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.982 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.979 \pm .016$	$.993 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.977 \pm .022$
	0.01	$.989 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .025$	$.974 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .014$
	0.05	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .005$	$.998 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .001$
	0.10	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 9: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human80 task expert and human60 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	**.848 ± .014**	.838 ± .019	.826 ± .019	.847 ± .008	.845 ± .006	.816 ± .014
	0.01	.834 ± .008	**.847 ± .013**	.825 ± .009	.823 ± .018	.841 ± .009	.819 ± .021
	0.05	.821 ± .016	**.842 ± .019**	.833 ± .008	.837 ± .021	.833 ± .014	.819 ± .014
	0.10	.828 ± .020	.812 ± .015	**.832 ± .014**	.827 ± .015	**.832 ± .024**	.829 ± .023
	0.15	.822 ± .004	.827 ± .022	.824 ± .015	**.839 ± .012**	.822 ± .012	.829 ± .008
	0.20	.821 ± .016	.824 ± .010	**.839 ± .020**	.826 ± .010	.831 ± .014	**.839 ± .014**
	0.25	.826 ± .002	**.833 ± .008**	.819 ± .016	.823 ± .015	.817 ± .018	.815 ± .023
	0.30	.828 ± .012	**.835 ± .014**	.828 ± .016	.812 ± .014	.822 ± .014	.819 ± .010
	0.35	.826 ± .012	.828 ± .012	.823 ± .021	**.842 ± .011**	.832 ± .015	.829 ± .013
	0.40	.828 ± .022	**.834 ± .016**	.825 ± .016	.830 ± .014	.832 ± .018	.831 ± .012
	0.45	.827 ± .014	.831 ± .014	.833 ± .024	.829 ± .009	**.837 ± .014**	.824 ± .018
	0.50	.808 ± .023	.823 ± .011	.821 ± .015	.806 ± .011	**.834 ± .010**	.814 ± .024
$AccConc$	0.00	.830 ± .009	**.845 ± .008**	.829 ± .007	.826 ± .007	.842 ± .008	.833 ± .010
	0.01	.827 ± .012	**.843 ± .010**	.831 ± .007	.831 ± .009	.843 ± .005	.827 ± .007
	0.05	.824 ± .006	**.845 ± .007**	.824 ± .012	.828 ± .011	.844 ± .003	.829 ± .008
	0.10	.833 ± .005	.843 ± .010	.832 ± .008	.822 ± .008	**.846 ± .006**	.834 ± .008
	0.15	.828 ± .008	.847 ± .008	.826 ± .009	.834 ± .008	**.847 ± .004**	.834 ± .009
	0.20	.823 ± .011	.844 ± .006	.830 ± .009	.826 ± .010	**.848 ± .007**	.831 ± .006
	0.25	.822 ± .008	.845 ± .003	.826 ± .008	.826 ± .007	**.845 ± .007**	.827 ± .015
	0.30	.827 ± .007	**.849 ± .008**	.826 ± .006	.821 ± .009	.841 ± .007	.824 ± .006
	0.35	.821 ± .011	**.845 ± .004**	.817 ± .010	.830 ± .006	.844 ± .012	.824 ± .006
	0.40	.821 ± .009	.842 ± .016	.824 ± .008	.825 ± .011	**.847 ± .007**	.815 ± .019
	0.45	.823 ± .010	**.845 ± .012**	.818 ± .017	.823 ± .006	.844 ± .012	.821 ± .013
	0.50	.812 ± .008	**.851 ± .007**	.816 ± .006	.814 ± .003	.841 ± .008	.817 ± .008
$CovTask$	0.00	.933 ± .021	.961 ± .020	**1.000 ± 0.000**	.949 ± .023	.943 ± .028	**1.000 ± 0.000**
	0.01	.969 ± .015	.943 ± .023	**1.000 ± 0.000**	.972 ± .021	.963 ± .022	**1.000 ± 0.000**
	0.05	.986 ± .012	.985 ± .019	**1.000 ± 0.000**	.977 ± .021	.996 ± .007	**1.000 ± 0.000**
	0.10	.992 ± .015	.999 ± .002	**1.000 ± 0.000**	.997 ± .007	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.15	**1.000 ± 0.000**	.997 ± .007	**1.000 ± 0.000**	.974 ± .045	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.20	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.25	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.30	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.35	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.40	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.45	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
$CovConc$	0.00	.973 ± .018	**1.000 ± 0.000**	.971 ± .017	.975 ± .013	**1.000 ± 0.000**	.979 ± .020
	0.01	.983 ± .016	**1.000 ± 0.000**	.971 ± .033	.976 ± .012	**1.000 ± 0.000**	.984 ± .003
	0.05	.991 ± .017	**1.000 ± 0.000**	.997 ± .004	.999 ± .002	**1.000 ± 0.000**	.999 ± .002
	0.10	**1.000 ± 0.000**	**1.000 ± 0.000**	.999 ± .002	1.000 ± .001	**1.000 ± 0.000**	1.000 ± 0.000
	0.15	**1.000 ± 0.000**	**1.000 ± 0.000**	1.000 ± 0.000	**1.000 ± 0.000**	**1.000 ± 0.000**	1.000 ± 0.000
	0.20	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.25	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.30	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.35	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.40	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.45	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**

Table 10: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human80 task expert and human60 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.841 \pm .020$	$.855 \pm .008$	$.840 \pm .010$	$.843 \pm .019$	$\mathbf{.856 \pm .022}$	$.827 \pm .019$
	0.01	$.839 \pm .016$	$.842 \pm .010$	$.815 \pm .008$	$.845 \pm .009$	$\mathbf{.851 \pm .013}$	$.820 \pm .006$
	0.05	$.842 \pm .008$	$\mathbf{.847 \pm .010}$	$.823 \pm .012$	$.832 \pm .021$	$.841 \pm .015$	$.827 \pm .017$
	0.10	$.832 \pm .011$	$.833 \pm .009$	$.814 \pm .011$	$.823 \pm .010$	$\mathbf{.834 \pm .011}$	$.824 \pm .022$
	0.15	$.815 \pm .008$	$.826 \pm .012$	$.829 \pm .007$	$.830 \pm .019$	$\mathbf{.831 \pm .011}$	$.828 \pm .010$
	0.20	$.823 \pm .015$	$\mathbf{.828 \pm .014}$	$.822 \pm .017$	$.827 \pm .015$	$\mathbf{.828 \pm .018}$	$.816 \pm .015$
	0.25	$.818 \pm .016$	$.833 \pm .023$	$.814 \pm .008$	$\mathbf{.837 \pm .012}$	$\mathbf{.837 \pm .012}$	$.831 \pm .022$
	0.30	$.830 \pm .011$	$.832 \pm .015$	$.827 \pm .014$	$\mathbf{.833 \pm .004}$	$.828 \pm .021$	$.822 \pm .008$
	0.35	$.826 \pm .019$	$.828 \pm .018$	$.826 \pm .009$	$.824 \pm .022$	$\mathbf{.837 \pm .004}$	$.836 \pm .008$
	0.40	$.819 \pm .019$	$\mathbf{.842 \pm .008}$	$.820 \pm .013$	$.836 \pm .011$	$.827 \pm .010$	$.823 \pm .007$
	0.45	$.810 \pm .018$	$.818 \pm .008$	$.817 \pm .025$	$\mathbf{.835 \pm .011}$	$.826 \pm .018$	$.825 \pm .021$
	0.50	$.815 \pm .038$	$.827 \pm .018$	$.809 \pm .014$	$.825 \pm .014$	$.827 \pm .012$	$\mathbf{.832 \pm .008}$
$AccConc$	0.00	$.834 \pm .010$	$\mathbf{.839 \pm .008}$	$.817 \pm .004$	$.829 \pm .007$	$.834 \pm .006$	$.821 \pm .011$
	0.01	$.823 \pm .009$	$.840 \pm .008$	$.822 \pm .007$	$.824 \pm .008$	$\mathbf{.844 \pm .009}$	$.830 \pm .011$
	0.05	$.827 \pm .004$	$.839 \pm .010$	$.824 \pm .012$	$.823 \pm .011$	$\mathbf{.840 \pm .010}$	$.827 \pm .011$
	0.10	$.830 \pm .008$	$\mathbf{.836 \pm .009}$	$.826 \pm .005$	$.827 \pm .008$	$.830 \pm .008$	$.820 \pm .003$
	0.15	$.829 \pm .006$	$.826 \pm .006$	$.834 \pm .014$	$.827 \pm .012$	$\mathbf{.834 \pm .012}$	$.830 \pm .009$
	0.20	$.817 \pm .008$	$.838 \pm .007$	$.824 \pm .018$	$.820 \pm .012$	$\mathbf{.844 \pm .010}$	$.825 \pm .013$
	0.25	$.832 \pm .008$	$.840 \pm .005$	$.823 \pm .007$	$.826 \pm .016$	$\mathbf{.845 \pm .006}$	$.819 \pm .012$
	0.30	$.822 \pm .008$	$\mathbf{.846 \pm .009}$	$.822 \pm .009$	$.823 \pm .007$	$.834 \pm .013$	$.823 \pm .012$
	0.35	$.815 \pm .010$	$\mathbf{.839 \pm .005}$	$.808 \pm .007$	$.821 \pm .006$	$.837 \pm .007$	$.815 \pm .006$
	0.40	$.810 \pm .004$	$\mathbf{.839 \pm .014}$	$.808 \pm .011$	$.823 \pm .008$	$.837 \pm .004$	$.808 \pm .017$
	0.45	$.811 \pm .007$	$.835 \pm .008$	$.817 \pm .011$	$.825 \pm .015$	$\mathbf{.839 \pm .007}$	$.818 \pm .009$
	0.50	$.813 \pm .016$	$\mathbf{.836 \pm .015}$	$.804 \pm .011$	$.814 \pm .006$	$.832 \pm .008$	$.811 \pm .007$
$CovTask$	0.00	$.794 \pm .217$	$.887 \pm .031$	$\mathbf{1.000 \pm 0.000}$	$.725 \pm .406$	$.908 \pm .047$	$\mathbf{1.000 \pm 0.000}$
	0.01	$.722 \pm .408$	$.793 \pm .268$	$\mathbf{1.000 \pm 0.000}$	$.921 \pm .044$	$.805 \pm .249$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.954 \pm .012$	$.948 \pm .012$	$\mathbf{1.000 \pm 0.000}$	$.826 \pm .229$	$.970 \pm .020$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.983 \pm .017$	$.992 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.992 \pm .009$	$.985 \pm .007$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.996 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.980 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.979 \pm .016$	$.991 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.977 \pm .022$
	0.01	$.990 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .025$	$.974 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .014$
	0.05	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .005$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .001$
	0.10	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 11: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering oracle task expert and human60 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.826 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.816 \pm .014$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.825 \pm .009$	$.999 \pm .002$	$.999 \pm .002$	$.819 \pm .021$
	0.05	$.952 \pm .028$	$.976 \pm .027$	$.833 \pm .008$	$.981 \pm .011$	$\mathbf{.986 \pm .012}$	$.819 \pm .014$
	0.10	$.898 \pm .021$	$.878 \pm .018$	$.832 \pm .014$	$.909 \pm .018$	$\mathbf{.921 \pm .038}$	$.829 \pm .023$
	0.15	$.872 \pm .009$	$\mathbf{.888 \pm .023}$	$.824 \pm .015$	$.877 \pm .011$	$.869 \pm .014$	$.829 \pm .008$
	0.20	$.844 \pm .019$	$.842 \pm .011$	$.839 \pm .020$	$.853 \pm .006$	$\mathbf{.862 \pm .015}$	$.839 \pm .014$
	0.25	$.826 \pm .004$	$\mathbf{.839 \pm .013}$	$.819 \pm .016$	$.826 \pm .013$	$.822 \pm .018$	$.815 \pm .023$
	0.30	$.828 \pm .008$	$\mathbf{.838 \pm .012}$	$.828 \pm .016$	$.827 \pm .014$	$.819 \pm .008$	$.819 \pm .010$
	0.35	$.821 \pm .015$	$.832 \pm .015$	$.823 \pm .021$	$\mathbf{.838 \pm .012}$	$.837 \pm .015$	$.829 \pm .013$
	0.40	$.828 \pm .020$	$\mathbf{.832 \pm .013}$	$.825 \pm .016$	$.831 \pm .013$	$.828 \pm .019$	$.831 \pm .012$
	0.45	$.826 \pm .012$	$\mathbf{.833 \pm .012}$	$\mathbf{.833 \pm .024}$	$.832 \pm .011$	$.830 \pm .015$	$.824 \pm .018$
	0.50	$.820 \pm .017$	$\mathbf{.831 \pm .010}$	$.821 \pm .015$	$.803 \pm .020$	$.830 \pm .009$	$.814 \pm .024$
$AccConc$	0.00	$.830 \pm .009$	$\mathbf{.845 \pm .008}$	$.829 \pm .007$	$.826 \pm .007$	$.842 \pm .008$	$.833 \pm .010$
	0.01	$.827 \pm .012$	$\mathbf{.843 \pm .010}$	$.831 \pm .007$	$.831 \pm .009$	$.843 \pm .005$	$.827 \pm .007$
	0.05	$.824 \pm .006$	$\mathbf{.845 \pm .007}$	$.824 \pm .012$	$.828 \pm .011$	$.844 \pm .003$	$.829 \pm .008$
	0.10	$.833 \pm .005$	$.843 \pm .010$	$.832 \pm .008$	$.822 \pm .008$	$\mathbf{.846 \pm .006}$	$.834 \pm .008$
	0.15	$.828 \pm .008$	$.847 \pm .008$	$.826 \pm .009$	$.834 \pm .008$	$\mathbf{.847 \pm .004}$	$.834 \pm .009$
	0.20	$.823 \pm .011$	$.844 \pm .006$	$.830 \pm .009$	$.826 \pm .010$	$\mathbf{.848 \pm .007}$	$.831 \pm .006$
	0.25	$.822 \pm .008$	$.845 \pm .003$	$.826 \pm .008$	$.826 \pm .007$	$\mathbf{.845 \pm .007}$	$.827 \pm .015$
	0.30	$.827 \pm .007$	$\mathbf{.849 \pm .008}$	$.826 \pm .006$	$.821 \pm .009$	$.841 \pm .007$	$.824 \pm .006$
	0.35	$.821 \pm .011$	$\mathbf{.845 \pm .004}$	$.817 \pm .010$	$.830 \pm .006$	$.844 \pm .012$	$.824 \pm .006$
	0.40	$.821 \pm .009$	$.842 \pm .016$	$.824 \pm .008$	$.825 \pm .011$	$\mathbf{.847 \pm .007}$	$.815 \pm .019$
	0.45	$.823 \pm .010$	$\mathbf{.845 \pm .012}$	$.818 \pm .017$	$.823 \pm .006$	$.844 \pm .012$	$.821 \pm .013$
	0.50	$.812 \pm .008$	$\mathbf{.851 \pm .007}$	$.816 \pm .006$	$.814 \pm .003$	$.841 \pm .008$	$.817 \pm .008$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	$.002 \pm .004$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .020$	$.005 \pm .011$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.351 \pm .173$	$.327 \pm .205$	$\mathbf{1.000 \pm 0.000}$	$.198 \pm .109$	$.159 \pm .103$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.793 \pm .123$	$.814 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.682 \pm .088$	$.684 \pm .279$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.893 \pm .031$	$.863 \pm .062$	$\mathbf{1.000 \pm 0.000}$	$.875 \pm .041$	$.896 \pm .020$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.949 \pm .023$	$.966 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.952 \pm .022$	$.941 \pm .025$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.994 \pm .005$	$.990 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.991 \pm .012$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.997 \pm .007$	$.995 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.973 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.971 \pm .017$	$.975 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.979 \pm .020$
	0.01	$.983 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.971 \pm .033$	$.976 \pm .012$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .003$
	0.05	$.991 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .004$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 12: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering oracle task expert and human60 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.840 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.827 \pm .019$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.815 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.820 \pm .006$
	0.05	$.991 \pm .008$	$.998 \pm .004$	$.823 \pm .012$	$.996 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.827 \pm .017$
	0.10	$\mathbf{.964 \pm .013}$	$.950 \pm .023$	$.814 \pm .011$	$.951 \pm .022$	$.957 \pm .029$	$.824 \pm .022$
	0.15	$.916 \pm .034$	$.903 \pm .024$	$.829 \pm .007$	$\mathbf{.944 \pm .039}$	$.899 \pm .014$	$.828 \pm .010$
	0.20	$\mathbf{.879 \pm .032}$	$.873 \pm .020$	$.822 \pm .017$	$.869 \pm .017$	$.878 \pm .010$	$.816 \pm .015$
	0.25	$.849 \pm .013$	$\mathbf{.854 \pm .013}$	$.814 \pm .008$	$.845 \pm .017$	$.848 \pm .016$	$.831 \pm .022$
	0.30	$.835 \pm .008$	$.837 \pm .012$	$.827 \pm .014$	$\mathbf{.844 \pm .005}$	$.839 \pm .024$	$.822 \pm .008$
	0.35	$.825 \pm .015$	$.831 \pm .016$	$.826 \pm .009$	$.816 \pm .011$	$.833 \pm .007$	$\mathbf{.836 \pm .008}$
	0.40	$.813 \pm .019$	$\mathbf{.834 \pm .016}$	$.820 \pm .013$	$.828 \pm .015$	$.827 \pm .006$	$.823 \pm .007$
	0.45	$.819 \pm .007$	$.822 \pm .009$	$.817 \pm .025$	$\mathbf{.826 \pm .009}$	$.825 \pm .023$	$.825 \pm .021$
	0.50	$.818 \pm .023$	$.821 \pm .020$	$.809 \pm .014$	$.823 \pm .013$	$.824 \pm .007$	$\mathbf{.832 \pm .008}$
$AccConc$	0.00	$.830 \pm .007$	$\mathbf{.839 \pm .009}$	$.817 \pm .004$	$.829 \pm .007$	$.834 \pm .007$	$.821 \pm .011$
	0.01	$.825 \pm .009$	$.841 \pm .008$	$.822 \pm .007$	$.820 \pm .007$	$\mathbf{.844 \pm .008}$	$.830 \pm .011$
	0.05	$.828 \pm .004$	$.839 \pm .007$	$.824 \pm .012$	$.824 \pm .009$	$\mathbf{.842 \pm .011}$	$.827 \pm .011$
	0.10	$.829 \pm .006$	$\mathbf{.839 \pm .004}$	$.826 \pm .005$	$.826 \pm .007$	$.834 \pm .006$	$.820 \pm .003$
	0.15	$.826 \pm .008$	$.830 \pm .008$	$.834 \pm .014$	$.823 \pm .013$	$\mathbf{.834 \pm .011}$	$.830 \pm .009$
	0.20	$.817 \pm .009$	$.839 \pm .006$	$.824 \pm .018$	$.820 \pm .010$	$\mathbf{.844 \pm .008}$	$.825 \pm .013$
	0.25	$.826 \pm .009$	$.843 \pm .006$	$.823 \pm .007$	$.826 \pm .013$	$\mathbf{.845 \pm .006}$	$.819 \pm .012$
	0.30	$.819 \pm .011$	$\mathbf{.844 \pm .012}$	$.822 \pm .009$	$.822 \pm .008$	$.834 \pm .010$	$.823 \pm .012$
	0.35	$.812 \pm .007$	$\mathbf{.838 \pm .003}$	$.808 \pm .007$	$.820 \pm .005$	$.838 \pm .006$	$.815 \pm .006$
	0.40	$.810 \pm .009$	$\mathbf{.840 \pm .012}$	$.808 \pm .011$	$.819 \pm .009$	$.838 \pm .007$	$.808 \pm .017$
	0.45	$.807 \pm .008$	$.834 \pm .008$	$.817 \pm .011$	$.826 \pm .011$	$\mathbf{.839 \pm .006}$	$.818 \pm .009$
	0.50	$.806 \pm .019$	$\mathbf{.834 \pm .009}$	$.804 \pm .011$	$.810 \pm .003$	$.832 \pm .007$	$.811 \pm .007$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.141 \pm .132$	$.086 \pm .090$	$\mathbf{1.000 \pm 0.000}$	$.040 \pm .089$	$.030 \pm .067$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.260 \pm .117$	$.478 \pm .139$	$\mathbf{1.000 \pm 0.000}$	$.420 \pm .174$	$.415 \pm .180$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.668 \pm .307$	$.801 \pm .069$	$\mathbf{1.000 \pm 0.000}$	$.463 \pm .340$	$.840 \pm .038$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.791 \pm .178$	$.893 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.894 \pm .057$	$.893 \pm .028$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.949 \pm .012$	$.955 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$.957 \pm .016$	$.961 \pm .017$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.994 \pm .007$	$.984 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .016$	$.994 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.979 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.979 \pm .016$	$.992 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.977 \pm .022$
	0.01	$.986 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .025$	$.974 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .014$
	0.05	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .005$	$.998 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .001$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 13: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human60 task expert and human80 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.822 \pm .016$	$.815 \pm .008$	$.820 \pm .009$	$.819 \pm .025$	$\mathbf{.828 \pm .010}$	$.821 \pm .013$
	0.01	$.823 \pm .019$	$.825 \pm .015$	$.817 \pm .023$	$.817 \pm .013$	$\mathbf{.831 \pm .017}$	$.824 \pm .007$
	0.05	$.821 \pm .024$	$.836 \pm .013$	$\mathbf{.838 \pm .024}$	$.816 \pm .012$	$.833 \pm .014$	$.827 \pm .010$
	0.10	$.823 \pm .008$	$.814 \pm .014$	$\mathbf{.835 \pm .006}$	$.823 \pm .012$	$.834 \pm .014$	$.830 \pm .023$
	0.15	$.824 \pm .016$	$\mathbf{.828 \pm .016}$	$.824 \pm .019$	$.822 \pm .015$	$.818 \pm .013$	$\mathbf{.828 \pm .010}$
	0.20	$.813 \pm .019$	$.825 \pm .014$	$.832 \pm .018$	$.825 \pm .009$	$.834 \pm .016$	$\mathbf{.836 \pm .015}$
	0.25	$.819 \pm .016$	$\mathbf{.836 \pm .011}$	$.824 \pm .020$	$.823 \pm .021$	$.817 \pm .019$	$.821 \pm .010$
	0.30	$.824 \pm .011$	$.831 \pm .016$	$\mathbf{.835 \pm .010}$	$.823 \pm .012$	$.821 \pm .010$	$.816 \pm .011$
	0.35	$.821 \pm .013$	$.822 \pm .014$	$.826 \pm .017$	$.830 \pm .017$	$\mathbf{.831 \pm .019}$	$.827 \pm .018$
	0.40	$.840 \pm .023$	$.824 \pm .019$	$.821 \pm .010$	$\mathbf{.841 \pm .011}$	$.834 \pm .016$	$.835 \pm .016$
	0.45	$.832 \pm .006$	$.835 \pm .014$	$\mathbf{.842 \pm .008}$	$.825 \pm .013$	$.827 \pm .016$	$.825 \pm .009$
	0.50	$.821 \pm .015$	$.818 \pm .012$	$.826 \pm .022$	$.801 \pm .012$	$\mathbf{.828 \pm .006}$	$.827 \pm .018$
$AccConc$	0.00	$.881 \pm .003$	$.845 \pm .008$	$.878 \pm .008$	$.879 \pm .005$	$.842 \pm .008$	$\mathbf{.883 \pm .008}$
	0.01	$.884 \pm .006$	$.843 \pm .010$	$.885 \pm .002$	$\mathbf{.885 \pm .005}$	$.843 \pm .005$	$.883 \pm .003$
	0.05	$.881 \pm .005$	$.845 \pm .007$	$\mathbf{.881 \pm .004}$	$.871 \pm .011$	$.844 \pm .003$	$.876 \pm .005$
	0.10	$.869 \pm .003$	$.843 \pm .010$	$.870 \pm .007$	$.864 \pm .003$	$.846 \pm .006$	$\mathbf{.873 \pm .007}$
	0.15	$.846 \pm .008$	$.847 \pm .008$	$.850 \pm .004$	$.850 \pm .007$	$.847 \pm .004$	$\mathbf{.853 \pm .003}$
	0.20	$.835 \pm .013$	$.844 \pm .006$	$.838 \pm .008$	$.837 \pm .006$	$\mathbf{.848 \pm .007}$	$.836 \pm .008$
	0.25	$.830 \pm .006$	$.845 \pm .003$	$.832 \pm .005$	$.833 \pm .008$	$\mathbf{.845 \pm .007}$	$.839 \pm .015$
	0.30	$.836 \pm .007$	$\mathbf{.849 \pm .008}$	$.834 \pm .004$	$.832 \pm .008$	$.841 \pm .007$	$.831 \pm .005$
	0.35	$.830 \pm .008$	$\mathbf{.845 \pm .004}$	$.824 \pm .013$	$.837 \pm .002$	$.844 \pm .012$	$.831 \pm .008$
	0.40	$.828 \pm .011$	$.842 \pm .016$	$.833 \pm .005$	$.832 \pm .012$	$\mathbf{.847 \pm .007}$	$.829 \pm .010$
	0.45	$.832 \pm .009$	$\mathbf{.845 \pm .012}$	$.829 \pm .011$	$.829 \pm .010$	$.844 \pm .012$	$.832 \pm .007$
	0.50	$.826 \pm .003$	$\mathbf{.851 \pm .007}$	$.821 \pm .008$	$.827 \pm .004$	$.841 \pm .008$	$.831 \pm .008$
$CovTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.438 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.412 \pm .045$	$.427 \pm .039$	$\mathbf{1.000 \pm 0.000}$	$.456 \pm .026$
	0.01	$.480 \pm .040$	$\mathbf{1.000 \pm 0.000}$	$.478 \pm .005$	$.478 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.480 \pm .011$
	0.05	$.603 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.595 \pm .040$	$.604 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.607 \pm .052$
	0.10	$.800 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.797 \pm .023$	$.773 \pm .043$	$\mathbf{1.000 \pm 0.000}$	$.783 \pm .021$
	0.15	$.933 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .023$	$.924 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .015$
	0.20	$.981 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .013$	$.983 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.980 \pm .008$
	0.25	$.997 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$.998 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .008$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 14: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human60 task expert and human80 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.810 \pm .015$	$.825 \pm .009$	$.834 \pm .013$	$.822 \pm .017$	$.824 \pm .014$	$\mathbf{.841 \pm .019}$
	0.01	$.825 \pm .015$	$\mathbf{.830 \pm .005}$	$.820 \pm .022$	$.825 \pm .009$	$.822 \pm .008$	$.826 \pm .012$
	0.05	$.819 \pm .002$	$.826 \pm .008$	$.831 \pm .023$	$.816 \pm .016$	$.832 \pm .014$	$\mathbf{.833 \pm .019}$
	0.10	$.825 \pm .014$	$.823 \pm .004$	$.822 \pm .011$	$.818 \pm .006$	$\mathbf{.830 \pm .007}$	$.825 \pm .009$
	0.15	$.809 \pm .016$	$.828 \pm .008$	$\mathbf{.833 \pm .006}$	$.823 \pm .017$	$.830 \pm .013$	$.829 \pm .009$
	0.20	$.816 \pm .014$	$.825 \pm .009$	$.823 \pm .014$	$.826 \pm .011$	$\mathbf{.830 \pm .015}$	$.826 \pm .015$
	0.25	$.816 \pm .002$	$\mathbf{.833 \pm .008}$	$.831 \pm .013$	$.833 \pm .014$	$.824 \pm .004$	$.828 \pm .025$
	0.30	$.822 \pm .010$	$.815 \pm .014$	$\mathbf{.828 \pm .016}$	$\mathbf{.828 \pm .009}$	$.827 \pm .010$	$.823 \pm .014$
	0.35	$.817 \pm .023$	$.828 \pm .021$	$.830 \pm .013$	$.833 \pm .018$	$.829 \pm .009$	$\mathbf{.837 \pm .013}$
	0.40	$.815 \pm .011$	$.823 \pm .008$	$\mathbf{.837 \pm .018}$	$.828 \pm .013$	$.833 \pm .018$	$.824 \pm .014$
	0.45	$.821 \pm .015$	$.815 \pm .006$	$.823 \pm .017$	$\mathbf{.833 \pm .009}$	$.825 \pm .015$	$.824 \pm .012$
	0.50	$.817 \pm .028$	$.818 \pm .017$	$.816 \pm .017$	$.827 \pm .006$	$\mathbf{.832 \pm .013}$	$.825 \pm .020$
$AccConc$	0.00	$\mathbf{.879 \pm .007}$	$.837 \pm .007$	$.876 \pm .005$	$.877 \pm .005$	$.832 \pm .007$	$.872 \pm .012$
	0.01	$\mathbf{.878 \pm .004}$	$.835 \pm .009$	$.869 \pm .003$	$.874 \pm .005$	$.842 \pm .014$	$.869 \pm .006$
	0.05	$.874 \pm .004$	$.837 \pm .011$	$\mathbf{.875 \pm .005}$	$.873 \pm .012$	$.838 \pm .005$	$.868 \pm .004$
	0.10	$.856 \pm .005$	$.839 \pm .007$	$\mathbf{.864 \pm .009}$	$.862 \pm .010$	$.830 \pm .008$	$.863 \pm .006$
	0.15	$.846 \pm .004$	$.827 \pm .006$	$.846 \pm .006$	$\mathbf{.848 \pm .007}$	$.834 \pm .011$	$.847 \pm .005$
	0.20	$.831 \pm .012$	$.839 \pm .011$	$.838 \pm .006$	$.833 \pm .006$	$\mathbf{.843 \pm .009}$	$.836 \pm .008$
	0.25	$.832 \pm .007$	$.839 \pm .004$	$.831 \pm .009$	$.834 \pm .015$	$\mathbf{.844 \pm .006}$	$.832 \pm .007$
	0.30	$.833 \pm .009$	$\mathbf{.844 \pm .008}$	$.826 \pm .009$	$.827 \pm .008$	$.831 \pm .009$	$.834 \pm .009$
	0.35	$.823 \pm .011$	$\mathbf{.838 \pm .006}$	$.825 \pm .003$	$.834 \pm .007$	$.834 \pm .005$	$.825 \pm .010$
	0.40	$.826 \pm .005$	$\mathbf{.841 \pm .011}$	$.819 \pm .010$	$.831 \pm .006$	$.838 \pm .003$	$.825 \pm .011$
	0.45	$.827 \pm .005$	$.837 \pm .010$	$.823 \pm .004$	$.835 \pm .006$	$\mathbf{.840 \pm .007}$	$.824 \pm .006$
	0.50	$.827 \pm .005$	$\mathbf{.835 \pm .013}$	$.820 \pm .009$	$.824 \pm .005$	$.835 \pm .006$	$.821 \pm .007$
$CovTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.491 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.490 \pm .043$	$.534 \pm .053$	$\mathbf{1.000 \pm 0.000}$	$.454 \pm .058$
	0.01	$.523 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.500 \pm .041$	$.507 \pm .031$	$\mathbf{1.000 \pm 0.000}$	$.487 \pm .033$
	0.05	$.653 \pm .052$	$\mathbf{1.000 \pm 0.000}$	$.631 \pm .028$	$.573 \pm .063$	$\mathbf{1.000 \pm 0.000}$	$.656 \pm .019$
	0.10	$.816 \pm .041$	$\mathbf{1.000 \pm 0.000}$	$.788 \pm .038$	$.838 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.805 \pm .049$
	0.15	$.919 \pm .047$	$\mathbf{1.000 \pm 0.000}$	$.953 \pm .023$	$.913 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.936 \pm .023$
	0.20	$.986 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .034$	$.967 \pm .041$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .022$
	0.25	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$.998 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 15: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human80 task expert and human80 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.846 \pm .023$	$.838 \pm .019$	$.820 \pm .009$	$\mathbf{.859 \pm .012}$	$.845 \pm .006$	$.821 \pm .013$
	0.01	$.839 \pm .019$	$\mathbf{.847 \pm .013}$	$.817 \pm .023$	$.840 \pm .015$	$.841 \pm .009$	$.824 \pm .007$
	0.05	$.829 \pm .014$	$\mathbf{.842 \pm .019}$	$.838 \pm .024$	$.832 \pm .019$	$.833 \pm .014$	$.827 \pm .010$
	0.10	$.822 \pm .004$	$.812 \pm .015$	$\mathbf{.835 \pm .006}$	$.827 \pm .010$	$.832 \pm .024$	$.830 \pm .023$
	0.15	$.820 \pm .015$	$.827 \pm .022$	$.824 \pm .019$	$\mathbf{.840 \pm .011}$	$.822 \pm .012$	$.828 \pm .010$
	0.20	$.826 \pm .012$	$.824 \pm .010$	$.832 \pm .018$	$.821 \pm .011$	$.831 \pm .014$	$\mathbf{.836 \pm .015}$
	0.25	$.823 \pm .018$	$\mathbf{.833 \pm .008}$	$.824 \pm .020$	$.825 \pm .019$	$.817 \pm .018$	$.821 \pm .010$
	0.30	$.819 \pm .011$	$\mathbf{.835 \pm .014}$	$\mathbf{.835 \pm .010}$	$.817 \pm .012$	$.822 \pm .014$	$.816 \pm .011$
	0.35	$.832 \pm .010$	$.828 \pm .012$	$.826 \pm .017$	$\mathbf{.837 \pm .017}$	$.832 \pm .015$	$.827 \pm .018$
	0.40	$.836 \pm .016$	$.834 \pm .016$	$.821 \pm .010$	$\mathbf{.846 \pm .014}$	$.832 \pm .018$	$.835 \pm .016$
	0.45	$.833 \pm .007$	$.831 \pm .014$	$\mathbf{.842 \pm .008}$	$.822 \pm .021$	$.837 \pm .014$	$.825 \pm .009$
	0.50	$.825 \pm .019$	$.823 \pm .011$	$.826 \pm .022$	$.805 \pm .015$	$\mathbf{.834 \pm .010}$	$.827 \pm .018$
$AccConc$	0.00	$.881 \pm .003$	$.845 \pm .008$	$.878 \pm .008$	$.879 \pm .005$	$.842 \pm .008$	$\mathbf{.883 \pm .008}$
	0.01	$.884 \pm .006$	$.843 \pm .010$	$.885 \pm .002$	$\mathbf{.885 \pm .005}$	$.843 \pm .005$	$.883 \pm .003$
	0.05	$.881 \pm .005$	$.845 \pm .007$	$\mathbf{.881 \pm .004}$	$.871 \pm .011$	$.844 \pm .003$	$.876 \pm .005$
	0.10	$.869 \pm .003$	$.843 \pm .010$	$.870 \pm .007$	$.864 \pm .003$	$.846 \pm .006$	$\mathbf{.873 \pm .007}$
	0.15	$.846 \pm .008$	$.847 \pm .008$	$.850 \pm .004$	$.850 \pm .007$	$.847 \pm .004$	$\mathbf{.853 \pm .003}$
	0.20	$.835 \pm .013$	$.844 \pm .006$	$.838 \pm .008$	$.837 \pm .006$	$\mathbf{.848 \pm .007}$	$.836 \pm .008$
	0.25	$.830 \pm .006$	$.845 \pm .003$	$.832 \pm .005$	$.833 \pm .008$	$\mathbf{.845 \pm .007}$	$.839 \pm .015$
	0.30	$.836 \pm .007$	$\mathbf{.849 \pm .008}$	$.834 \pm .004$	$.832 \pm .008$	$.841 \pm .007$	$.831 \pm .005$
	0.35	$.830 \pm .008$	$\mathbf{.845 \pm .004}$	$.824 \pm .013$	$.837 \pm .002$	$.844 \pm .012$	$.831 \pm .008$
	0.40	$.828 \pm .011$	$.842 \pm .016$	$.833 \pm .005$	$.832 \pm .012$	$\mathbf{.847 \pm .007}$	$.829 \pm .010$
	0.45	$.832 \pm .009$	$\mathbf{.845 \pm .012}$	$.829 \pm .011$	$.829 \pm .010$	$.844 \pm .012$	$.832 \pm .007$
	0.50	$.826 \pm .003$	$\mathbf{.851 \pm .007}$	$.821 \pm .008$	$.827 \pm .004$	$.841 \pm .008$	$.831 \pm .008$
$CovTask$	0.00	$.901 \pm .023$	$.961 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.901 \pm .027$	$.943 \pm .028$	$\mathbf{1.000 \pm 0.000}$
	0.01	$.940 \pm .025$	$.943 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.933 \pm .039$	$.963 \pm .022$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.982 \pm .012$	$.985 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.961 \pm .028$	$.996 \pm .007$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.993 \pm .011$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.993 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.971 \pm .057$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.438 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.412 \pm .045$	$.427 \pm .039$	$\mathbf{1.000 \pm 0.000}$	$.456 \pm .026$
	0.01	$.480 \pm .040$	$\mathbf{1.000 \pm 0.000}$	$.478 \pm .005$	$.478 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.480 \pm .011$
	0.05	$.603 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.595 \pm .040$	$.604 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.607 \pm .052$
	0.10	$.800 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.797 \pm .023$	$.773 \pm .043$	$\mathbf{1.000 \pm 0.000}$	$.783 \pm .021$
	0.15	$.933 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .023$	$.924 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .015$
	0.20	$.981 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .013$	$.983 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.980 \pm .008$
	0.25	$.997 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$.998 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .008$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 16: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human80 task expert and human80 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.833 \pm .019$	$.841 \pm .004$	$.834 \pm .013$	$\mathbf{.853 \pm .015}$	$.846 \pm .019$	$.841 \pm .019$
	0.01	$.850 \pm .018$	$.845 \pm .010$	$.820 \pm .022$	$\mathbf{.853 \pm .010}$	$.851 \pm .008$	$.826 \pm .012$
	0.05	$.831 \pm .009$	$.841 \pm .013$	$.831 \pm .023$	$\mathbf{.849 \pm .016}$	$.838 \pm .014$	$.833 \pm .019$
	0.10	$\mathbf{.828 \pm .020}$	$.822 \pm .014$	$.822 \pm .011$	$.820 \pm .011$	$.827 \pm .008$	$.825 \pm .009$
	0.15	$.812 \pm .013$	$.828 \pm .008$	$.833 \pm .006$	$.824 \pm .011$	$\mathbf{.834 \pm .010}$	$.829 \pm .009$
	0.20	$.816 \pm .015$	$.821 \pm .009$	$.823 \pm .014$	$.816 \pm .007$	$\mathbf{.827 \pm .010}$	$.826 \pm .015$
	0.25	$.820 \pm .010$	$\mathbf{.838 \pm .006}$	$.831 \pm .013$	$.832 \pm .014$	$.826 \pm .013$	$.828 \pm .025$
	0.30	$.826 \pm .007$	$.829 \pm .004$	$.828 \pm .016$	$\mathbf{.832 \pm .010}$	$.819 \pm .015$	$.823 \pm .014$
	0.35	$.830 \pm .022$	$.832 \pm .021$	$.830 \pm .013$	$\mathbf{.838 \pm .010}$	$.836 \pm .013$	$.837 \pm .013$
	0.40	$.828 \pm .018$	$.830 \pm .012$	$\mathbf{.837 \pm .018}$	$.831 \pm .015$	$.833 \pm .008$	$.824 \pm .014$
	0.45	$.820 \pm .016$	$.819 \pm .007$	$.823 \pm .017$	$\mathbf{.835 \pm .011}$	$\mathbf{.835 \pm .016}$	$.824 \pm .012$
	0.50	$.824 \pm .026$	$.820 \pm .018$	$.816 \pm .017$	$.826 \pm .011$	$\mathbf{.831 \pm .007}$	$.825 \pm .020$
$AccConc$	0.00	$\mathbf{.881 \pm .007}$	$.839 \pm .007$	$.876 \pm .005$	$.877 \pm .005$	$.834 \pm .008$	$.872 \pm .012$
	0.01	$\mathbf{.879 \pm .005}$	$.839 \pm .008$	$.869 \pm .003$	$.875 \pm .004$	$.844 \pm .011$	$.869 \pm .006$
	0.05	$.876 \pm .004$	$.840 \pm .007$	$.875 \pm .005$	$\mathbf{.877 \pm .011}$	$.841 \pm .008$	$.868 \pm .004$
	0.10	$.858 \pm .004$	$.838 \pm .006$	$\mathbf{.864 \pm .009}$	$.862 \pm .007$	$.829 \pm .007$	$.863 \pm .006$
	0.15	$.846 \pm .002$	$.827 \pm .006$	$.846 \pm .006$	$\mathbf{.848 \pm .009}$	$.834 \pm .009$	$.847 \pm .005$
	0.20	$.830 \pm .013$	$.839 \pm .009$	$.838 \pm .006$	$.832 \pm .007$	$\mathbf{.844 \pm .011}$	$.836 \pm .008$
	0.25	$.831 \pm .008$	$.838 \pm .004$	$.831 \pm .009$	$.834 \pm .014$	$\mathbf{.845 \pm .006}$	$.832 \pm .007$
	0.30	$.832 \pm .011$	$\mathbf{.845 \pm .008}$	$.826 \pm .009$	$.827 \pm .006$	$.836 \pm .007$	$.834 \pm .009$
	0.35	$.824 \pm .012$	$\mathbf{.836 \pm .006}$	$.825 \pm .003$	$.832 \pm .009$	$.835 \pm .009$	$.825 \pm .010$
	0.40	$.825 \pm .002$	$\mathbf{.840 \pm .013}$	$.819 \pm .010$	$.829 \pm .006$	$.837 \pm .004$	$.825 \pm .011$
	0.45	$.827 \pm .002$	$.835 \pm .007$	$.823 \pm .004$	$.831 \pm .008$	$\mathbf{.837 \pm .006}$	$.824 \pm .006$
	0.50	$.829 \pm .007$	$\mathbf{.836 \pm .013}$	$.820 \pm .009$	$.825 \pm .005$	$.836 \pm .007$	$.821 \pm .007$
$CovTask$	0.00	$.804 \pm .231$	$.922 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.903 \pm .020$	$.936 \pm .037$	$\mathbf{1.000 \pm 0.000}$
	0.01	$.888 \pm .060$	$.845 \pm .188$	$\mathbf{1.000 \pm 0.000}$	$.905 \pm .015$	$.916 \pm .025$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.967 \pm .009$	$.968 \pm .012$	$\mathbf{1.000 \pm 0.000}$	$.938 \pm .030$	$.968 \pm .012$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.989 \pm .009$	$.992 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.993 \pm .010$	$.996 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.500 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.490 \pm .043$	$.536 \pm .055$	$\mathbf{1.000 \pm 0.000}$	$.454 \pm .058$
	0.01	$.526 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.500 \pm .041$	$.498 \pm .029$	$\mathbf{1.000 \pm 0.000}$	$.487 \pm .033$
	0.05	$.649 \pm .045$	$\mathbf{1.000 \pm 0.000}$	$.631 \pm .028$	$.583 \pm .050$	$\mathbf{1.000 \pm 0.000}$	$.656 \pm .019$
	0.10	$.817 \pm .046$	$\mathbf{1.000 \pm 0.000}$	$.788 \pm .038$	$.834 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.805 \pm .049$
	0.15	$.916 \pm .046$	$\mathbf{1.000 \pm 0.000}$	$.953 \pm .023$	$.911 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.936 \pm .023$
	0.20	$.987 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .034$	$.970 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .022$
	0.25	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 17: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering oracle task expert and human80 concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.820 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.821 \pm .013$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.817 \pm .023$	$.999 \pm .002$	$.999 \pm .002$	$.824 \pm .007$
	0.05	$.967 \pm .019$	$.976 \pm .027$	$.838 \pm .024$	$.975 \pm .014$	$\mathbf{.986 \pm .012}$	$.827 \pm .010$
	0.10	$.904 \pm .022$	$.878 \pm .018$	$.835 \pm .006$	$.912 \pm .010$	$\mathbf{.921 \pm .038}$	$.830 \pm .023$
	0.15	$.871 \pm .013$	$\mathbf{.888 \pm .023}$	$.824 \pm .019$	$.879 \pm .014$	$.869 \pm .014$	$.828 \pm .010$
	0.20	$.846 \pm .015$	$.842 \pm .011$	$.832 \pm .018$	$.845 \pm .009$	$\mathbf{.862 \pm .015}$	$.836 \pm .015$
	0.25	$.826 \pm .014$	$\mathbf{.839 \pm .013}$	$.824 \pm .020$	$.828 \pm .014$	$.822 \pm .018$	$.821 \pm .010$
	0.30	$.819 \pm .013$	$\mathbf{.838 \pm .012}$	$.835 \pm .010$	$.829 \pm .010$	$.819 \pm .008$	$.816 \pm .011$
	0.35	$.831 \pm .009$	$.832 \pm .015$	$.826 \pm .017$	$.835 \pm .018$	$\mathbf{.837 \pm .015}$	$.827 \pm .018$
	0.40	$.839 \pm .015$	$.832 \pm .013$	$.821 \pm .010$	$\mathbf{.844 \pm .012}$	$.828 \pm .019$	$.835 \pm .016$
	0.45	$.834 \pm .010$	$.833 \pm .012$	$\mathbf{.842 \pm .008}$	$.824 \pm .021$	$.830 \pm .015$	$.825 \pm .009$
	0.50	$.823 \pm .012$	$\mathbf{.831 \pm .010}$	$.826 \pm .022$	$.805 \pm .017$	$.830 \pm .009$	$.827 \pm .018$
$AccConc$	0.00	$.881 \pm .003$	$.845 \pm .008$	$.878 \pm .008$	$.879 \pm .005$	$.842 \pm .008$	$\mathbf{.883 \pm .008}$
	0.01	$.884 \pm .006$	$.843 \pm .010$	$.885 \pm .002$	$\mathbf{.885 \pm .005}$	$.843 \pm .005$	$.883 \pm .003$
	0.05	$.881 \pm .005$	$.845 \pm .007$	$\mathbf{.881 \pm .004}$	$.871 \pm .011$	$.844 \pm .003$	$.876 \pm .005$
	0.10	$.869 \pm .003$	$.843 \pm .010$	$.870 \pm .007$	$.864 \pm .003$	$.846 \pm .006$	$\mathbf{.873 \pm .007}$
	0.15	$.846 \pm .008$	$.847 \pm .008$	$.850 \pm .004$	$.850 \pm .007$	$.847 \pm .004$	$\mathbf{.853 \pm .003}$
	0.20	$.835 \pm .013$	$.844 \pm .006$	$.838 \pm .008$	$.837 \pm .006$	$\mathbf{.848 \pm .007}$	$.836 \pm .008$
	0.25	$.830 \pm .006$	$.845 \pm .003$	$.832 \pm .005$	$.833 \pm .008$	$\mathbf{.845 \pm .007}$	$.839 \pm .015$
	0.30	$.836 \pm .007$	$\mathbf{.849 \pm .008}$	$.834 \pm .004$	$.832 \pm .008$	$.841 \pm .007$	$.831 \pm .005$
	0.35	$.830 \pm .008$	$\mathbf{.845 \pm .004}$	$.824 \pm .013$	$.837 \pm .002$	$.844 \pm .012$	$.831 \pm .008$
	0.40	$.828 \pm .011$	$.842 \pm .016$	$.833 \pm .005$	$.832 \pm .012$	$\mathbf{.847 \pm .007}$	$.829 \pm .010$
	0.45	$.832 \pm .009$	$\mathbf{.845 \pm .012}$	$.829 \pm .011$	$.829 \pm .010$	$.844 \pm .012$	$.832 \pm .007$
	0.50	$.826 \pm .003$	$\mathbf{.851 \pm .007}$	$.821 \pm .008$	$.827 \pm .004$	$.841 \pm .008$	$.831 \pm .008$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	$.005 \pm .007$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .017$	$.005 \pm .011$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.332 \pm .140$	$.327 \pm .205$	$\mathbf{1.000 \pm 0.000}$	$.205 \pm .117$	$.159 \pm .103$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.768 \pm .127$	$.814 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.670 \pm .092$	$.684 \pm .279$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.891 \pm .015$	$.863 \pm .062$	$\mathbf{1.000 \pm 0.000}$	$.870 \pm .041$	$.896 \pm .020$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.947 \pm .022$	$.966 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.949 \pm .029$	$.941 \pm .025$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.988 \pm .010$	$.990 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.991 \pm .013$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.997 \pm .007$	$.995 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.987 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.438 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.412 \pm .045$	$.427 \pm .039$	$\mathbf{1.000 \pm 0.000}$	$.456 \pm .026$
	0.01	$.480 \pm .040$	$\mathbf{1.000 \pm 0.000}$	$.478 \pm .005$	$.478 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.480 \pm .011$
	0.05	$.603 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.595 \pm .040$	$.604 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.607 \pm .052$
	0.10	$.800 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.797 \pm .023$	$.773 \pm .043$	$\mathbf{1.000 \pm 0.000}$	$.783 \pm .021$
	0.15	$.933 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .023$	$.924 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.908 \pm .015$
	0.20	$.981 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .013$	$.983 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.980 \pm .008$
	0.25	$.997 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$.998 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .008$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 18: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering oracle task expert and human80 concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.834 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.841 \pm .019$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.820 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.826 \pm .012$
	0.05	$.988 \pm .008$	$.981 \pm .012$	$.831 \pm .023$	$\mathbf{.992 \pm .012}$	$.987 \pm .018$	$.833 \pm .019$
	0.10	$\mathbf{.946 \pm .010}$	$.943 \pm .025$	$.822 \pm .011$	$.927 \pm .021$	$.926 \pm .025$	$.825 \pm .009$
	0.15	$.886 \pm .010$	$\mathbf{.887 \pm .018}$	$.833 \pm .006$	$.885 \pm .017$	$.882 \pm .027$	$.829 \pm .009$
	0.20	$.854 \pm .019$	$\mathbf{.869 \pm .011}$	$.823 \pm .014$	$.854 \pm .016$	$.864 \pm .012$	$.826 \pm .015$
	0.25	$.844 \pm .013$	$\mathbf{.853 \pm .008}$	$.831 \pm .013$	$.850 \pm .016$	$.852 \pm .010$	$.828 \pm .025$
	0.30	$.820 \pm .015$	$.839 \pm .018$	$.828 \pm .016$	$\mathbf{.843 \pm .014}$	$.834 \pm .013$	$.823 \pm .014$
	0.35	$\mathbf{.843 \pm .020}$	$.838 \pm .010$	$.830 \pm .013$	$.842 \pm .012$	$.837 \pm .010$	$.837 \pm .013$
	0.40	$.817 \pm .012$	$\mathbf{.841 \pm .007}$	$.837 \pm .018$	$.831 \pm .013$	$.833 \pm .013$	$.824 \pm .014$
	0.45	$.826 \pm .005$	$.819 \pm .007$	$.823 \pm .017$	$\mathbf{.837 \pm .014}$	$.837 \pm .016$	$.824 \pm .012$
	0.50	$\mathbf{.830 \pm .018}$	$.823 \pm .017$	$.816 \pm .017$	$.829 \pm .002$	$.824 \pm .009$	$.825 \pm .020$
$AccConc$	0.00	$\mathbf{.883 \pm .008}$	$.839 \pm .010$	$.876 \pm .005$	$.877 \pm .006$	$.836 \pm .006$	$.872 \pm .012$
	0.01	$\mathbf{.878 \pm .006}$	$.840 \pm .007$	$.869 \pm .003$	$.876 \pm .004$	$.842 \pm .009$	$.869 \pm .006$
	0.05	$.874 \pm .004$	$.840 \pm .006$	$.875 \pm .005$	$\mathbf{.875 \pm .013}$	$.843 \pm .008$	$.868 \pm .004$
	0.10	$.858 \pm .007$	$.839 \pm .008$	$\mathbf{.864 \pm .009}$	$.862 \pm .010$	$.834 \pm .005$	$.863 \pm .006$
	0.15	$.846 \pm .006$	$.830 \pm .007$	$.846 \pm .006$	$\mathbf{.848 \pm .010}$	$.835 \pm .014$	$.847 \pm .005$
	0.20	$.830 \pm .011$	$.836 \pm .005$	$.838 \pm .006$	$.834 \pm .006$	$\mathbf{.845 \pm .009}$	$.836 \pm .008$
	0.25	$.831 \pm .008$	$.841 \pm .005$	$.831 \pm .009$	$.835 \pm .015$	$\mathbf{.846 \pm .006}$	$.832 \pm .007$
	0.30	$.831 \pm .011$	$\mathbf{.844 \pm .008}$	$.826 \pm .009$	$.829 \pm .005$	$.839 \pm .008$	$.834 \pm .009$
	0.35	$.822 \pm .009$	$.835 \pm .005$	$.825 \pm .003$	$.832 \pm .009$	$\mathbf{.836 \pm .006}$	$.825 \pm .010$
	0.40	$.823 \pm .006$	$\mathbf{.841 \pm .009}$	$.819 \pm .010$	$.829 \pm .006$	$.837 \pm .004$	$.825 \pm .011$
	0.45	$.824 \pm .005$	$.836 \pm .007$	$.823 \pm .004$	$.835 \pm .008$	$\mathbf{.839 \pm .005}$	$.824 \pm .006$
	0.50	$.825 \pm .001$	$\mathbf{.837 \pm .012}$	$.820 \pm .009$	$.823 \pm .003$	$.834 \pm .007$	$.821 \pm .007$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.129 \pm .088$	$.246 \pm .138$	$\mathbf{1.000 \pm 0.000}$	$.071 \pm .066$	$.167 \pm .164$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.436 \pm .136$	$.576 \pm .188$	$\mathbf{1.000 \pm 0.000}$	$.585 \pm .161$	$.637 \pm .113$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.862 \pm .026$	$.865 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.833 \pm .088$	$.872 \pm .048$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.918 \pm .018$	$.909 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.932 \pm .024$	$.929 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.959 \pm .019$	$.965 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$.973 \pm .013$	$.962 \pm .023$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$.989 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.986 \pm .016$	$.986 \pm .014$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.994 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.522 \pm .030$	$\mathbf{1.000 \pm 0.000}$	$.490 \pm .043$	$.549 \pm .051$	$\mathbf{1.000 \pm 0.000}$	$.454 \pm .058$
	0.01	$.549 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$.500 \pm .041$	$.524 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.487 \pm .033$
	0.05	$.667 \pm .044$	$\mathbf{1.000 \pm 0.000}$	$.631 \pm .028$	$.604 \pm .053$	$\mathbf{1.000 \pm 0.000}$	$.656 \pm .019$
	0.10	$.827 \pm .042$	$\mathbf{1.000 \pm 0.000}$	$.788 \pm .038$	$.843 \pm .040$	$\mathbf{1.000 \pm 0.000}$	$.805 \pm .049$
	0.15	$.926 \pm .045$	$\mathbf{1.000 \pm 0.000}$	$.953 \pm .023$	$.916 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.936 \pm .023$
	0.20	$.990 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.981 \pm .034$	$.972 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .022$
	0.25	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 19: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human60 task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.881 \pm .020$	$.815 \pm .008$	$.906 \pm .032$	$.884 \pm .032$	$.828 \pm .010$	$\mathbf{.925 \pm .010}$
	0.01	$.917 \pm .021$	$.825 \pm .015$	$.914 \pm .019$	$.887 \pm .026$	$.831 \pm .017$	$\mathbf{.919 \pm .022}$
	0.05	$.893 \pm .019$	$.836 \pm .013$	$\mathbf{.914 \pm .020}$	$.884 \pm .029$	$.833 \pm .014$	$.904 \pm .025$
	0.10	$.885 \pm .033$	$.814 \pm .014$	$\mathbf{.921 \pm .020}$	$.882 \pm .033$	$.834 \pm .014$	$.905 \pm .034$
	0.15	$.873 \pm .038$	$.828 \pm .016$	$\mathbf{.913 \pm .025}$	$.866 \pm .035$	$.818 \pm .013$	$.910 \pm .006$
	0.20	$.864 \pm .029$	$.825 \pm .014$	$.884 \pm .007$	$.852 \pm .014$	$.834 \pm .016$	$\mathbf{.886 \pm .002}$
	0.25	$.845 \pm .027$	$.836 \pm .011$	$.851 \pm .031$	$.829 \pm .018$	$.817 \pm .019$	$\mathbf{.871 \pm .025}$
	0.30	$\mathbf{.838 \pm .016}$	$.831 \pm .016$	$.823 \pm .019$	$.831 \pm .018$	$.821 \pm .010$	$.830 \pm .018$
	0.35	$.810 \pm .015$	$.822 \pm .014$	$.833 \pm .016$	$\mathbf{.834 \pm .027}$	$.831 \pm .019$	$.830 \pm .017$
	0.40	$.834 \pm .017$	$.824 \pm .019$	$.831 \pm .013$	$.834 \pm .009$	$.834 \pm .016$	$\mathbf{.839 \pm .011}$
	0.45	$.836 \pm .015$	$.835 \pm .014$	$\mathbf{.837 \pm .011}$	$.834 \pm .004$	$.827 \pm .016$	$.826 \pm .018$
	0.50	$.822 \pm .018$	$.818 \pm .012$	$\mathbf{.834 \pm .011}$	$.807 \pm .010$	$.828 \pm .006$	$.832 \pm .013$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.845 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.843 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.843 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{.999 \pm .001}$	$.845 \pm .007$	$.999 \pm .001$	$.999 \pm .001$	$.844 \pm .003$	$.998 \pm .001$
	0.10	$.995 \pm .001$	$.843 \pm .010$	$.996 \pm .001$	$.996 \pm .001$	$.846 \pm .006$	$\mathbf{.996 \pm .001}$
	0.15	$.988 \pm .002$	$.847 \pm .008$	$\mathbf{.992 \pm .001}$	$.988 \pm .001$	$.847 \pm .004$	$.989 \pm .001$
	0.20	$.975 \pm .006$	$.844 \pm .006$	$.975 \pm .002$	$.973 \pm .003$	$.848 \pm .007$	$\mathbf{.977 \pm .002}$
	0.25	$.942 \pm .005$	$.845 \pm .003$	$\mathbf{.944 \pm .004}$	$.940 \pm .007$	$.845 \pm .007$	$.943 \pm .006$
	0.30	$.899 \pm .008$	$.849 \pm .008$	$.892 \pm .006$	$\mathbf{.901 \pm .006}$	$.841 \pm .007$	$.898 \pm .006$
	0.35	$.864 \pm .005$	$.845 \pm .004$	$.863 \pm .005$	$\mathbf{.867 \pm .002}$	$.844 \pm .012$	$.861 \pm .004$
	0.40	$.843 \pm .008$	$.842 \pm .016$	$.846 \pm .002$	$.846 \pm .010$	$.847 \pm .007$	$\mathbf{.848 \pm .008}$
	0.45	$\mathbf{.849 \pm .008}$	$.845 \pm .012$	$.842 \pm .010$	$.839 \pm .010$	$.844 \pm .012$	$.837 \pm .005$
	0.50	$.840 \pm .004$	$\mathbf{.851 \pm .007}$	$.836 \pm .008$	$.839 \pm .005$	$.841 \pm .008$	$.840 \pm .006$
$CovTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$0.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.001 \pm .001$
	0.01	$.011 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .005$	$.019 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.007 \pm .004$
	0.05	$.082 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.096 \pm .028$	$.089 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.107 \pm .019$
	0.10	$.241 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.225 \pm .025$	$.215 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.214 \pm .030$
	0.15	$.356 \pm .041$	$\mathbf{1.000 \pm 0.000}$	$.330 \pm .031$	$.375 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.363 \pm .046$
	0.20	$.484 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.496 \pm .031$	$.519 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.485 \pm .025$
	0.25	$.638 \pm .031$	$\mathbf{1.000 \pm 0.000}$	$.652 \pm .019$	$.645 \pm .046$	$\mathbf{1.000 \pm 0.000}$	$.670 \pm .012$
	0.30	$.800 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .016$	$.815 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.824 \pm .030$
	0.35	$.943 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.920 \pm .009$	$.928 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.935 \pm .019$
	0.40	$.994 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.994 \pm .005$	$.993 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.988 \pm .010$
	0.45	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 20: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human60 task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.887 \pm .031$	$.825 \pm .009$	$.906 \pm .026$	$.899 \pm .033$	$.828 \pm .012$	$\mathbf{.924 \pm .026}$
	0.01	$.886 \pm .019$	$.829 \pm .012$	$\mathbf{.930 \pm .013}$	$.883 \pm .014$	$.830 \pm .010$	$.909 \pm .024$
	0.05	$.888 \pm .027$	$.836 \pm .014$	$.896 \pm .023$	$.888 \pm .021$	$.834 \pm .015$	$\mathbf{.927 \pm .003}$
	0.10	$.902 \pm .027$	$.810 \pm .015$	$.903 \pm .029$	$.866 \pm .009$	$.823 \pm .009$	$\mathbf{.919 \pm .005}$
	0.15	$.867 \pm .006$	$.817 \pm .022$	$.898 \pm .006$	$.885 \pm .023$	$.818 \pm .017$	$\mathbf{.913 \pm .018}$
	0.20	$.863 \pm .020$	$.825 \pm .019$	$.889 \pm .005$	$.877 \pm .028$	$.832 \pm .008$	$\mathbf{.892 \pm .025}$
	0.25	$.859 \pm .028$	$.835 \pm .009$	$.878 \pm .018$	$.862 \pm .006$	$.825 \pm .015$	$\mathbf{.882 \pm .016}$
	0.30	$.845 \pm .017$	$.832 \pm .008$	$.849 \pm .012$	$.853 \pm .011$	$.832 \pm .010$	$\mathbf{.855 \pm .023}$
	0.35	$.838 \pm .019$	$.825 \pm .015$	$.850 \pm .008$	$.831 \pm .013$	$.834 \pm .018$	$\mathbf{.853 \pm .023}$
	0.40	$.825 \pm .006$	$.825 \pm .015$	$\mathbf{.840 \pm .005}$	$.830 \pm .015$	$.826 \pm .011$	$.835 \pm .012$
	0.45	$.833 \pm .013$	$.822 \pm .010$	$.831 \pm .016$	$\mathbf{.843 \pm .014}$	$.833 \pm .012$	$.833 \pm .013$
	0.50	$\mathbf{.840 \pm .014}$	$.825 \pm .010$	$.826 \pm .008$	$.837 \pm .009$	$.820 \pm .017$	$.829 \pm .012$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.839 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.838 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.840 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.844 \pm .010$	$\mathbf{1.000 \pm 0.000}$
	0.05	1.000 ± 0.000	$.840 \pm .009$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .005$	$.999 \pm .001$
	0.10	$.997 \pm .001$	$.840 \pm .004$	$\mathbf{.997 \pm .001}$	$.997 \pm .001$	$.834 \pm .005$	$.997 \pm .001$
	0.15	$\mathbf{.991 \pm .003}$	$.831 \pm .005$	$.987 \pm .004$	$.989 \pm .005$	$.836 \pm .010$	$.989 \pm .004$
	0.20	$.973 \pm .007$	$.840 \pm .008$	$.975 \pm .002$	$.976 \pm .004$	$.845 \pm .008$	$\mathbf{.977 \pm .003}$
	0.25	$.945 \pm .006$	$.843 \pm .007$	$.948 \pm .006$	$\mathbf{.949 \pm .008}$	$.846 \pm .003$	$.946 \pm .007$
	0.30	$\mathbf{.915 \pm .003}$	$.846 \pm .006$	$.903 \pm .006$	$.910 \pm .007$	$.835 \pm .009$	$.905 \pm .008$
	0.35	$.858 \pm .007$	$.837 \pm .010$	$.861 \pm .009$	$\mathbf{.869 \pm .004}$	$.836 \pm .004$	$.865 \pm .007$
	0.40	$.837 \pm .007$	$.840 \pm .009$	$.841 \pm .008$	$\mathbf{.846 \pm .010}$	$.838 \pm .006$	$.839 \pm .004$
	0.45	$.836 \pm .003$	$.838 \pm .008$	$.833 \pm .008$	$\mathbf{.843 \pm .005}$	$.842 \pm .006$	$.834 \pm .007$
	0.50	$.836 \pm .006$	$\mathbf{.837 \pm .013}$	$.835 \pm .008$	$.835 \pm .007$	$.834 \pm .007$	$.832 \pm .007$
$CovTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.10	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.001 \pm .002$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$0.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$0.000 \pm .001$
	0.01	$.009 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .009$	$.005 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .006$
	0.05	$.086 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .028$	$.059 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .005$
	0.10	$.170 \pm .040$	$\mathbf{1.000 \pm 0.000}$	$.174 \pm .023$	$.188 \pm .029$	$\mathbf{1.000 \pm 0.000}$	$.193 \pm .027$
	0.15	$.306 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.322 \pm .036$	$.347 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$.344 \pm .012$
	0.20	$.471 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.474 \pm .018$	$.442 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.449 \pm .026$
	0.25	$.628 \pm .045$	$\mathbf{1.000 \pm 0.000}$	$.599 \pm .034$	$.618 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.628 \pm .038$
	0.30	$.745 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.773 \pm .015$	$.767 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.791 \pm .023$
	0.35	$.927 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.927 \pm .018$	$.916 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.916 \pm .029$
	0.40	$.982 \pm .012$	$\mathbf{1.000 \pm 0.000}$	$.970 \pm .008$	$.965 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.978 \pm .014$
	0.45	$.998 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$	$.999 \pm 0.000$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$
	0.50	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000

Table 21: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering human80 task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	.921 ± .021	.838 ± .019	.906 ± .032	.919 ± .019	.845 ± .006	**.925 ± .010**
	0.01	.902 ± .017	.847 ± .013	.914 ± .019	.903 ± .030	.841 ± .009	**.919 ± .022**
	0.05	.905 ± .019	.842 ± .019	**.914 ± .020**	.897 ± .033	.833 ± .014	.904 ± .025
	0.10	.877 ± .035	.812 ± .015	**.921 ± .020**	.890 ± .034	.832 ± .024	.905 ± .034
	0.15	.863 ± .016	.827 ± .022	**.913 ± .025**	.873 ± .030	.822 ± .012	.910 ± .006
	0.20	.866 ± .026	.824 ± .010	.884 ± .007	.850 ± .013	.831 ± .014	**.886 ± .002**
	0.25	.853 ± .028	.833 ± .008	.851 ± .031	.833 ± .018	.817 ± .018	**.871 ± .025**
	0.30	**.842 ± .019**	.835 ± .014	.823 ± .019	.831 ± .020	.822 ± .014	.830 ± .018
	0.35	.825 ± .016	.828 ± .012	.833 ± .016	**.837 ± .020**	.832 ± .015	.830 ± .017
	0.40	.835 ± .019	.834 ± .016	.831 ± .013	.838 ± .010	.832 ± .018	**.839 ± .011**
	0.45	**.839 ± .014**	.831 ± .014	.837 ± .011	.834 ± .007	.837 ± .014	.826 ± .018
	0.50	.821 ± .014	.823 ± .011	**.834 ± .011**	.805 ± .013	**.834 ± .010**	.832 ± .013
$AccConc$	0.00	**1.000 ± 0.000**	.845 ± .008	**1.000 ± 0.000**	**1.000 ± 0.000**	.842 ± .008	**1.000 ± 0.000**
	0.01	**1.000 ± 0.000**	.843 ± .010	**1.000 ± 0.000**	**1.000 ± 0.000**	.843 ± .005	**1.000 ± 0.000**
	0.05	**.999 ± .001**	.845 ± .007	.999 ± .001	.999 ± .001	.844 ± .003	.998 ± .001
	0.10	.995 ± .001	.843 ± .010	.996 ± .001	.996 ± .001	.846 ± .006	**.996 ± .001**
	0.15	.988 ± .002	.847 ± .008	**.992 ± .001**	.988 ± .001	.847 ± .004	.989 ± .001
	0.20	.975 ± .006	.844 ± .006	.975 ± .002	.973 ± .003	.848 ± .007	**.977 ± .002**
	0.25	.942 ± .005	.845 ± .003	**.944 ± .004**	.940 ± .007	.845 ± .007	.943 ± .006
	0.30	.899 ± .008	.849 ± .008	.892 ± .006	**.901 ± .006**	.841 ± .007	.898 ± .006
	0.35	.864 ± .005	.845 ± .004	.863 ± .005	**.867 ± .002**	.844 ± .012	.861 ± .004
	0.40	.843 ± .008	.842 ± .016	.846 ± .002	.846 ± .010	.847 ± .007	**.848 ± .008**
	0.45	**.849 ± .008**	.845 ± .012	.842 ± .010	.839 ± .010	.844 ± .012	.837 ± .005
	0.50	.840 ± .004	**.851 ± .007**	.836 ± .008	.839 ± .005	.841 ± .008	.840 ± .006
$CovTask$	0.00	.917 ± .020	.961 ± .020	**1.000 ± 0.000**	.922 ± .015	.943 ± .028	**1.000 ± 0.000**
	0.01	.949 ± .014	.943 ± .023	**1.000 ± 0.000**	.946 ± .046	.963 ± .022	**1.000 ± 0.000**
	0.05	.974 ± .022	.985 ± .019	**1.000 ± 0.000**	.959 ± .039	.996 ± .007	**1.000 ± 0.000**
	0.10	.991 ± .013	.999 ± .002	**1.000 ± 0.000**	.987 ± .029	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.15	.999 ± .002	.997 ± .007	**1.000 ± 0.000**	.975 ± .050	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.20	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.25	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.30	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.35	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.40	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.45	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**
$CovConc$	0.00	0.000 ± .001	**1.000 ± 0.000**	0.000 ± 0.000	.001 ± .001	**1.000 ± 0.000**	.001 ± .001
	0.01	.011 ± .003	**1.000 ± 0.000**	.009 ± .005	.019 ± .010	**1.000 ± 0.000**	.007 ± .004
	0.05	.082 ± .021	**1.000 ± 0.000**	.096 ± .028	.089 ± .026	**1.000 ± 0.000**	.107 ± .019
	0.10	.241 ± .035	**1.000 ± 0.000**	.225 ± .025	.215 ± .027	**1.000 ± 0.000**	.214 ± .030
	0.15	.356 ± .041	**1.000 ± 0.000**	.330 ± .031	.375 ± .024	**1.000 ± 0.000**	.363 ± .046
	0.20	.484 ± .027	**1.000 ± 0.000**	.496 ± .031	.519 ± .027	**1.000 ± 0.000**	.485 ± .025
	0.25	.638 ± .031	**1.000 ± 0.000**	.652 ± .019	.645 ± .046	**1.000 ± 0.000**	.670 ± .012
	0.30	.800 ± .032	**1.000 ± 0.000**	.842 ± .016	.815 ± .028	**1.000 ± 0.000**	.824 ± .030
	0.35	.943 ± .008	**1.000 ± 0.000**	.920 ± .009	.928 ± .017	**1.000 ± 0.000**	.935 ± .019
	0.40	.994 ± .004	**1.000 ± 0.000**	.994 ± .005	.993 ± .007	**1.000 ± 0.000**	.988 ± .010
	0.45	1.000 ± 0.000	**1.000 ± 0.000**	**1.000 ± 0.000**	1.000 ± .001	**1.000 ± 0.000**	.999 ± .001
	0.50	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**	**1.000 ± 0.000**

Table 22: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering human80 task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$.892 \pm .024$	$.838 \pm .010$	$.906 \pm .026$	$\mathbf{.930 \pm .004}$	$.848 \pm .010$	$.924 \pm .026$
	0.01	$.903 \pm .028$	$.846 \pm .016$	$\mathbf{.930 \pm .013}$	$.902 \pm .023$	$.848 \pm .006$	$.909 \pm .024$
	0.05	$.897 \pm .022$	$.830 \pm .021$	$.896 \pm .023$	$.891 \pm .019$	$.839 \pm .027$	$\mathbf{.927 \pm .003}$
	0.10	$.890 \pm .020$	$.820 \pm .017$	$.903 \pm .029$	$.868 \pm .007$	$.827 \pm .013$	$\mathbf{.919 \pm .005}$
	0.15	$.877 \pm .034$	$.822 \pm .021$	$.898 \pm .006$	$.888 \pm .026$	$.824 \pm .019$	$\mathbf{.913 \pm .018}$
	0.20	$.868 \pm .016$	$.832 \pm .019$	$.889 \pm .005$	$\mathbf{.903 \pm .018}$	$.837 \pm .021$	$.892 \pm .025$
	0.25	$.866 \pm .021$	$.843 \pm .011$	$.878 \pm .018$	$.868 \pm .020$	$.819 \pm .016$	$\mathbf{.882 \pm .016}$
	0.30	$.850 \pm .011$	$.830 \pm .007$	$.849 \pm .012$	$\mathbf{.859 \pm .023}$	$.833 \pm .013$	$.855 \pm .023$
	0.35	$.836 \pm .011$	$.827 \pm .015$	$.850 \pm .008$	$.841 \pm .013$	$.835 \pm .017$	$\mathbf{.853 \pm .023}$
	0.40	$.834 \pm .020$	$.830 \pm .010$	$\mathbf{.840 \pm .005}$	$.831 \pm .016$	$.827 \pm .014$	$.835 \pm .012$
	0.45	$.834 \pm .010$	$.819 \pm .011$	$.831 \pm .016$	$\mathbf{.838 \pm .014}$	$.835 \pm .009$	$.833 \pm .013$
	0.50	$\mathbf{.835 \pm .008}$	$.827 \pm .009$	$.826 \pm .008$	$\mathbf{.835 \pm .008}$	$.824 \pm .009$	$.829 \pm .012$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.840 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.835 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.840 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.845 \pm .012$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.999 \pm 0.000$	$.841 \pm .010$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.844 \pm .006$	$.999 \pm .001$
	0.10	$.996 \pm .001$	$.841 \pm .004$	$\mathbf{.997 \pm .001}$	$.997 \pm .001$	$.833 \pm .006$	$.997 \pm .001$
	0.15	$\mathbf{.991 \pm .002}$	$.829 \pm .007$	$.987 \pm .004$	$.990 \pm .002$	$.835 \pm .009$	$.989 \pm .004$
	0.20	$.973 \pm .006$	$.840 \pm .009$	$.975 \pm .002$	$.975 \pm .006$	$.846 \pm .009$	$\mathbf{.977 \pm .003}$
	0.25	$.948 \pm .007$	$.843 \pm .006$	$.948 \pm .006$	$\mathbf{.949 \pm .008}$	$.846 \pm .002$	$.946 \pm .007$
	0.30	$\mathbf{.914 \pm .006}$	$.846 \pm .007$	$.903 \pm .006$	$.907 \pm .006$	$.840 \pm .009$	$.905 \pm .008$
	0.35	$.860 \pm .008$	$.834 \pm .009$	$.861 \pm .009$	$\mathbf{.867 \pm .005}$	$.835 \pm .005$	$.865 \pm .007$
	0.40	$.836 \pm .007$	$.840 \pm .009$	$.841 \pm .008$	$\mathbf{.846 \pm .009}$	$.838 \pm .006$	$.839 \pm .004$
	0.45	$.834 \pm .003$	$.836 \pm .007$	$.833 \pm .008$	$.842 \pm .004$	$\mathbf{.842 \pm .008}$	$.834 \pm .007$
	0.50	$.837 \pm .007$	$\mathbf{.838 \pm .012}$	$.835 \pm .008$	$.832 \pm .007$	$.836 \pm .006$	$.832 \pm .007$
$CovTask$	0.00	$.838 \pm .210$	$.943 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.919 \pm .031$	$.939 \pm .022$	$\mathbf{1.000 \pm 0.000}$
	0.01	$.930 \pm .024$	$.953 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.946 \pm .041$	$.942 \pm .017$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.973 \pm .030$	$.989 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.983 \pm .014$	$.986 \pm .019$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.999 \pm .002$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.15	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.20	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.25	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$0.000 \pm .001$
	0.01	$.012 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .009$	$.005 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .006$
	0.05	$.086 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .028$	$.061 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .005$
	0.10	$.170 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.174 \pm .023$	$.190 \pm .029$	$\mathbf{1.000 \pm 0.000}$	$.193 \pm .027$
	0.15	$.304 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.322 \pm .036$	$.341 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.344 \pm .012$
	0.20	$.470 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$.474 \pm .018$	$.445 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.449 \pm .026$
	0.25	$.620 \pm .049$	$\mathbf{1.000 \pm 0.000}$	$.599 \pm .034$	$.616 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.628 \pm .038$
	0.30	$.747 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.773 \pm .015$	$.774 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.791 \pm .023$
	0.35	$.924 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.927 \pm .018$	$.919 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.916 \pm .029$
	0.40	$.985 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.970 \pm .008$	$.964 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.978 \pm .014$
	0.45	$.998 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$
	0.50	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000

Table 23: Results for the `completeness` dataset when not allowing for shared parameters with independent training using ASM, and considering oracle task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.906 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.925 \pm .010$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.914 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$.919 \pm .022$
	0.05	$\mathbf{.999 \pm .002}$	$.976 \pm .027$	$.914 \pm .020$	$.997 \pm .004$	$.986 \pm .012$	$.904 \pm .025$
	0.10	$.979 \pm .004$	$.878 \pm .018$	$.921 \pm .020$	$\mathbf{.982 \pm .008}$	$.921 \pm .038$	$.905 \pm .034$
	0.15	$.938 \pm .023$	$.888 \pm .023$	$.913 \pm .025$	$\mathbf{.945 \pm .012}$	$.869 \pm .014$	$.910 \pm .006$
	0.20	$\mathbf{.890 \pm .010}$	$.842 \pm .011$	$.884 \pm .007$	$\mathbf{.890 \pm .008}$	$.862 \pm .015$	$.886 \pm .002$
	0.25	$.859 \pm .030$	$.839 \pm .013$	$.851 \pm .031$	$.853 \pm .008$	$.822 \pm .018$	$\mathbf{.871 \pm .025}$
	0.30	$.841 \pm .008$	$.838 \pm .012$	$.823 \pm .019$	$\mathbf{.842 \pm .018}$	$.819 \pm .008$	$.830 \pm .018$
	0.35	$.827 \pm .016$	$.832 \pm .015$	$.833 \pm .016$	$.835 \pm .018$	$\mathbf{.837 \pm .015}$	$.830 \pm .017$
	0.40	$.832 \pm .013$	$.832 \pm .013$	$.831 \pm .013$	$\mathbf{.842 \pm .014}$	$.828 \pm .019$	$.839 \pm .011$
	0.45	$\mathbf{.842 \pm .015}$	$.833 \pm .012$	$.837 \pm .011$	$.831 \pm .007$	$.830 \pm .015$	$.826 \pm .018$
	0.50	$.833 \pm .010$	$.831 \pm .010$	$\mathbf{.834 \pm .011}$	$.808 \pm .017$	$.830 \pm .009$	$.832 \pm .013$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.845 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.843 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.843 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{.999 \pm .001}$	$.845 \pm .007$	$.999 \pm .001$	$.999 \pm .001$	$.844 \pm .003$	$.998 \pm .001$
	0.10	$.995 \pm .001$	$.843 \pm .010$	$.996 \pm .001$	$.996 \pm .001$	$.846 \pm .006$	$\mathbf{.996 \pm .001}$
	0.15	$.988 \pm .002$	$.847 \pm .008$	$\mathbf{.992 \pm .001}$	$.988 \pm .001$	$.847 \pm .004$	$.989 \pm .001$
	0.20	$.975 \pm .006$	$.844 \pm .006$	$.975 \pm .002$	$.973 \pm .003$	$.848 \pm .007$	$\mathbf{.977 \pm .002}$
	0.25	$.942 \pm .005$	$.845 \pm .003$	$\mathbf{.944 \pm .004}$	$.940 \pm .007$	$.845 \pm .007$	$.943 \pm .006$
	0.30	$.899 \pm .008$	$.849 \pm .008$	$.892 \pm .006$	$\mathbf{.901 \pm .006}$	$.841 \pm .007$	$.898 \pm .006$
	0.35	$.864 \pm .005$	$.845 \pm .004$	$.863 \pm .005$	$\mathbf{.867 \pm .002}$	$.844 \pm .012$	$.861 \pm .004$
	0.40	$.843 \pm .008$	$.842 \pm .016$	$.846 \pm .002$	$.846 \pm .010$	$.847 \pm .007$	$\mathbf{.848 \pm .008}$
	0.45	$\mathbf{.849 \pm .008}$	$.845 \pm .012$	$.842 \pm .010$	$.839 \pm .010$	$.844 \pm .012$	$.837 \pm .005$
	0.50	$.840 \pm .004$	$\mathbf{.851 \pm .007}$	$.836 \pm .008$	$.839 \pm .005$	$.841 \pm .008$	$.840 \pm .006$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	$.004 \pm .007$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.008 \pm .018$	$.005 \pm .011$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.305 \pm .134$	$.327 \pm .205$	$\mathbf{1.000 \pm 0.000}$	$.212 \pm .108$	$.159 \pm .103$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.659 \pm .116$	$.814 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.601 \pm .056$	$.684 \pm .279$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.865 \pm .030$	$.863 \pm .062$	$\mathbf{1.000 \pm 0.000}$	$.818 \pm .059$	$.896 \pm .020$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.944 \pm .015$	$.966 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.934 \pm .030$	$.941 \pm .025$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.984 \pm .014$	$.990 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.980 \pm .024$	$.998 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.992 \pm .013$	$.995 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.988 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$0.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.001 \pm .001$
	0.01	$.011 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .005$	$.019 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.007 \pm .004$
	0.05	$.082 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.096 \pm .028$	$.089 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.107 \pm .019$
	0.10	$.241 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.225 \pm .025$	$.215 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.214 \pm .030$
	0.15	$.356 \pm .041$	$\mathbf{1.000 \pm 0.000}$	$.330 \pm .031$	$.375 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.363 \pm .046$
	0.20	$.484 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.496 \pm .031$	$.519 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.485 \pm .025$
	0.25	$.638 \pm .031$	$\mathbf{1.000 \pm 0.000}$	$.652 \pm .019$	$.645 \pm .046$	$\mathbf{1.000 \pm 0.000}$	$.670 \pm .012$
	0.30	$.800 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .016$	$.815 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.824 \pm .030$
	0.35	$.943 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.920 \pm .009$	$.928 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.935 \pm .019$
	0.40	$.994 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.994 \pm .005$	$.993 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.988 \pm .010$
	0.45	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$1.000 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .001$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 24: Results for the `completeness` dataset when not allowing for shared parameters with joint training using ASM, and considering oracle task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.906 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.924 \pm .026$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.930 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.909 \pm .024$
	0.05	$\mathbf{1.000 \pm 0.000}$	$.979 \pm .013$	$.896 \pm .023$	$.997 \pm .004$	$.986 \pm .009$	$.927 \pm .003$
	0.10	$\mathbf{.979 \pm .008}$	$.900 \pm .026$	$.903 \pm .029$	$.974 \pm .007$	$.900 \pm .015$	$.919 \pm .005$
	0.15	$.948 \pm .010$	$.869 \pm .015$	$.898 \pm .006$	$\mathbf{.949 \pm .011}$	$.861 \pm .017$	$.913 \pm .018$
	0.20	$.892 \pm .022$	$.856 \pm .026$	$.889 \pm .005$	$\mathbf{.920 \pm .024}$	$.876 \pm .013$	$.892 \pm .025$
	0.25	$.862 \pm .020$	$.845 \pm .011$	$.878 \pm .018$	$.878 \pm .020$	$.836 \pm .012$	$\mathbf{.882 \pm .016}$
	0.30	$.859 \pm .018$	$.834 \pm .010$	$.849 \pm .012$	$\mathbf{.871 \pm .031}$	$.830 \pm .008$	$.855 \pm .023$
	0.35	$.839 \pm .008$	$.833 \pm .013$	$.850 \pm .008$	$.848 \pm .021$	$.845 \pm .022$	$\mathbf{.853 \pm .023}$
	0.40	$.833 \pm .016$	$.829 \pm .010$	$.840 \pm .005$	$.828 \pm .018$	$\mathbf{.841 \pm .015}$	$.835 \pm .012$
	0.45	$\mathbf{.840 \pm .020}$	$.822 \pm .010$	$.831 \pm .016$	$.839 \pm .011$	$.833 \pm .009$	$.833 \pm .013$
	0.50	$.840 \pm .016$	$.824 \pm .014$	$.826 \pm .008$	$\mathbf{.845 \pm .005}$	$.824 \pm .011$	$.829 \pm .012$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.839 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.837 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.841 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.844 \pm .012$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.999 \pm .001$	$.844 \pm .008$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.846 \pm .005$	$.999 \pm .001$
	0.10	$.996 \pm .001$	$.838 \pm .007$	$\mathbf{.997 \pm .001}$	$.996 \pm .001$	$.834 \pm .005$	$.997 \pm .001$
	0.15	$\mathbf{.991 \pm .003}$	$.833 \pm .009$	$.987 \pm .004$	$.991 \pm .002$	$.836 \pm .008$	$.989 \pm .004$
	0.20	$.972 \pm .007$	$.841 \pm .008$	$.975 \pm .002$	$.975 \pm .004$	$.848 \pm .010$	$\mathbf{.977 \pm .003}$
	0.25	$.945 \pm .006$	$.842 \pm .006$	$\mathbf{.948 \pm .006}$	$.948 \pm .008$	$.846 \pm .003$	$.946 \pm .007$
	0.30	$\mathbf{.912 \pm .004}$	$.845 \pm .005$	$.903 \pm .006$	$.908 \pm .006$	$.838 \pm .008$	$.905 \pm .008$
	0.35	$.860 \pm .007$	$.833 \pm .007$	$.861 \pm .009$	$\mathbf{.868 \pm .004}$	$.837 \pm .005$	$.865 \pm .007$
	0.40	$.837 \pm .008$	$.839 \pm .010$	$.841 \pm .008$	$\mathbf{.846 \pm .012}$	$.836 \pm .006$	$.839 \pm .004$
	0.45	$.833 \pm .001$	$.837 \pm .005$	$.833 \pm .008$	$.840 \pm .005$	$\mathbf{.842 \pm .008}$	$.834 \pm .007$
	0.50	$.836 \pm .005$	$\mathbf{.837 \pm .012}$	$.835 \pm .008$	$.832 \pm .008$	$.836 \pm .010$	$.832 \pm .007$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	$.014 \pm .031$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.003 \pm .007$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.215 \pm .093$	$.268 \pm .103$	$\mathbf{1.000 \pm 0.000}$	$.136 \pm .108$	$.178 \pm .099$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.609 \pm .132$	$.764 \pm .076$	$\mathbf{1.000 \pm 0.000}$	$.671 \pm .030$	$.771 \pm .058$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.808 \pm .050$	$.905 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.740 \pm .142$	$.903 \pm .038$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.936 \pm .027$	$.937 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.902 \pm .065$	$.921 \pm .031$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.972 \pm .013$	$.980 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.982 \pm .025$	$.978 \pm .017$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.977 \pm .044$	$.996 \pm .005$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$0.000 \pm .001$
	0.01	$.011 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .009$	$.006 \pm .005$	$\mathbf{1.000 \pm 0.000}$	$.009 \pm .006$
	0.05	$.090 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .028$	$.065 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.083 \pm .005$
	0.10	$.173 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.174 \pm .023$	$.197 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.193 \pm .027$
	0.15	$.311 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.322 \pm .036$	$.349 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.344 \pm .012$
	0.20	$.473 \pm .030$	$\mathbf{1.000 \pm 0.000}$	$.474 \pm .018$	$.451 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.449 \pm .026$
	0.25	$.630 \pm .049$	$\mathbf{1.000 \pm 0.000}$	$.599 \pm .034$	$.620 \pm .030$	$\mathbf{1.000 \pm 0.000}$	$.628 \pm .038$
	0.30	$.748 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.773 \pm .015$	$.774 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.791 \pm .023$
	0.35	$.912 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.927 \pm .018$	$.916 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.916 \pm .029$
	0.40	$.977 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.970 \pm .008$	$.961 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.978 \pm .014$
	0.45	$.997 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$	$.999 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .004$
	0.50	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000

Table 25: Results for the `completeness` dataset when allowing for shared parameters with independent training using ASM, and considering oracle task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.895 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.907 \pm .023$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.914 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.935 \pm .004$
	0.05	$\mathbf{.996 \pm .004}$	$.974 \pm .011$	$.922 \pm .015$	$.992 \pm .006$	$.958 \pm .032$	$.923 \pm .008$
	0.10	$\mathbf{.978 \pm .004}$	$.926 \pm .047$	$.923 \pm .010$	$.969 \pm .022$	$.914 \pm .014$	$.901 \pm .024$
	0.15	$\mathbf{.945 \pm .014}$	$.885 \pm .032$	$.908 \pm .010$	$\mathbf{.945 \pm .007}$	$.866 \pm .016$	$.911 \pm .023$
	0.20	$\mathbf{.900 \pm .021}$	$.850 \pm .013$	$.858 \pm .008$	$.891 \pm .024$	$.860 \pm .010$	$.876 \pm .009$
	0.25	$.847 \pm .024$	$.846 \pm .016$	$\mathbf{.852 \pm .016}$	$.850 \pm .018$	$.830 \pm .015$	$.841 \pm .017$
	0.30	$\mathbf{.839 \pm .024}$	$.831 \pm .010$	$.838 \pm .023$	$.826 \pm .010$	$.834 \pm .020$	$.833 \pm .011$
	0.35	$.823 \pm .010$	$.827 \pm .018$	$.829 \pm .011$	$.833 \pm .010$	$\mathbf{.843 \pm .013}$	$.832 \pm .008$
	0.40	$.823 \pm .018$	$.827 \pm .014$	$.824 \pm .012$	$\mathbf{.830 \pm .011}$	$.829 \pm .008$	$.826 \pm .016$
	0.45	$.823 \pm .014$	$.827 \pm .016$	$.822 \pm .004$	$.821 \pm .014$	$\mathbf{.838 \pm .015}$	$.828 \pm .012$
	0.50	$.819 \pm .009$	$.831 \pm .016$	$.823 \pm .014$	$.821 \pm .009$	$.829 \pm .007$	$\mathbf{.832 \pm .008}$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.863 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.869 \pm .006$	$\mathbf{1.000 \pm 0.000}$
	0.01	1.000 ± 0.000	$.869 \pm .007$	1.000 ± 0.000	1.000 ± 0.000	$.871 \pm .012$	$\mathbf{1.000 \pm 0.000}$
	0.05	$\mathbf{.998 \pm .001}$	$.870 \pm .004$	$.998 \pm .001$	$.997 \pm .001$	$.872 \pm .009$	$.998 \pm .001$
	0.10	$.994 \pm .001$	$.867 \pm .006$	$\mathbf{.995 \pm .002}$	$.994 \pm .002$	$.866 \pm .014$	$.994 \pm .001$
	0.15	$\mathbf{.988 \pm .001}$	$.865 \pm .010$	$.988 \pm .003$	$.985 \pm .002$	$.872 \pm .008$	$.988 \pm .001$
	0.20	$.974 \pm .002$	$.863 \pm .010$	$.973 \pm .003$	$.970 \pm .001$	$.862 \pm .011$	$\mathbf{.975 \pm .004}$
	0.25	$.941 \pm .008$	$.866 \pm .008$	$.948 \pm .009$	$\mathbf{.949 \pm .003}$	$.869 \pm .009$	$.940 \pm .004$
	0.30	$\mathbf{.911 \pm .006}$	$.861 \pm .008$	$.911 \pm .004$	$.905 \pm .007$	$.869 \pm .006$	$.905 \pm .007$
	0.35	$.883 \pm .009$	$.869 \pm .008$	$.882 \pm .004$	$\mathbf{.887 \pm .001}$	$.871 \pm .008$	$.874 \pm .010$
	0.40	$.863 \pm .010$	$.858 \pm .009$	$.868 \pm .011$	$.865 \pm .012$	$\mathbf{.872 \pm .006}$	$.869 \pm .007$
	0.45	$.863 \pm .007$	$.860 \pm .014$	$\mathbf{.872 \pm .014}$	$.870 \pm .010$	$.872 \pm .010$	$.861 \pm .010$
	0.50	$.860 \pm .008$	$.866 \pm .012$	$\mathbf{.869 \pm .005}$	$.862 \pm .013$	$.868 \pm .012$	$.864 \pm .008$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	$.015 \pm .034$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.257 \pm .172$	$.287 \pm .151$	$\mathbf{1.000 \pm 0.000}$	$.265 \pm .132$	$.360 \pm .253$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.693 \pm .064$	$.607 \pm .349$	$\mathbf{1.000 \pm 0.000}$	$.638 \pm .277$	$.659 \pm .155$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.819 \pm .067$	$.807 \pm .182$	$\mathbf{1.000 \pm 0.000}$	$.860 \pm .017$	$.924 \pm .026$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.931 \pm .025$	$.963 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.953 \pm .024$	$.931 \pm .022$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.990 \pm .010$	$.980 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.994 \pm .008$	$.993 \pm .006$	$\mathbf{1.000 \pm 0.000}$
	0.30	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.35	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .007$	$.997 \pm .007$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.001 \pm .002$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000
	0.01	$.021 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.019 \pm .007$	$.027 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.023 \pm .009$
	0.05	$.144 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.123 \pm .055$	$.162 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.157 \pm .030$
	0.10	$.295 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.288 \pm .022$	$.270 \pm .041$	$\mathbf{1.000 \pm 0.000}$	$.299 \pm .014$
	0.15	$.406 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.410 \pm .011$	$.438 \pm .030$	$\mathbf{1.000 \pm 0.000}$	$.434 \pm .015$
	0.20	$.567 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.562 \pm .014$	$.561 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.539 \pm .043$
	0.25	$.712 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.706 \pm .051$	$.710 \pm .027$	$\mathbf{1.000 \pm 0.000}$	$.717 \pm .022$
	0.30	$.869 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.856 \pm .034$	$.869 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.854 \pm .015$
	0.35	$.955 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.957 \pm .012$	$.959 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.958 \pm .014$
	0.40	$.993 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .003$	$.994 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .004$
	0.45	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000
	0.50	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 26: Results for the `completeness` dataset when allowing for shared parameters with joint training using ASM, and considering oracle task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.933 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.919 \pm .017$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.922 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.925 \pm .020$
	0.05	$\mathbf{1.000 \pm 0.000}$	$.963 \pm .036$	$.905 \pm .045$	$.995 \pm .004$	$.957 \pm .010$	$.920 \pm .009$
	0.10	$.963 \pm .028$	$.898 \pm .038$	$.900 \pm .032$	$\mathbf{.981 \pm .007}$	$.914 \pm .022$	$.925 \pm .010$
	0.15	$\mathbf{.945 \pm .011}$	$.876 \pm .021$	$.919 \pm .011$	$.937 \pm .013$	$.865 \pm .006$	$.893 \pm .031$
	0.20	$.901 \pm .027$	$.848 \pm .014$	$.890 \pm .020$	$\mathbf{.911 \pm .025}$	$.840 \pm .017$	$.871 \pm .028$
	0.25	$.854 \pm .035$	$.837 \pm .008$	$.866 \pm .017$	$\mathbf{.868 \pm .014}$	$.845 \pm .028$	$.863 \pm .029$
	0.30	$.844 \pm .009$	$.828 \pm .012$	$.843 \pm .020$	$\mathbf{.855 \pm .009}$	$.830 \pm .013$	$.844 \pm .016$
	0.35	$.833 \pm .011$	$.822 \pm .016$	$.836 \pm .024$	$\mathbf{.840 \pm .013}$	$.822 \pm .009$	$.830 \pm .018$
	0.40	$.831 \pm .007$	$.833 \pm .006$	$.821 \pm .018$	$\mathbf{.838 \pm .012}$	$.812 \pm .012$	$.828 \pm .017$
	0.45	$.829 \pm .009$	$.825 \pm .017$	$.832 \pm .009$	$.832 \pm .014$	$.822 \pm .003$	$\mathbf{.833 \pm .014}$
	0.50	$.818 \pm .024$	$.816 \pm .015$	$.828 \pm .018$	$\mathbf{.836 \pm .014}$	$.823 \pm .013$	$.829 \pm .025$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.866 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.856 \pm .010$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.859 \pm .005$	1.000 ± 0.000	1.000 ± 0.000	$.869 \pm .004$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.997 \pm .001$	$.861 \pm .009$	$.999 \pm .001$	$.998 \pm .001$	$.853 \pm .008$	$\mathbf{.999 \pm .001}$
	0.10	$.993 \pm .004$	$.852 \pm .010$	$\mathbf{.996 \pm .001}$	$.994 \pm .001$	$.857 \pm .015$	$.995 \pm .001$
	0.15	$.983 \pm .006$	$.864 \pm .015$	$.987 \pm .006$	$.984 \pm .003$	$.862 \pm .011$	$\mathbf{.989 \pm .003}$
	0.20	$.967 \pm .008$	$.859 \pm .007$	$.972 \pm .006$	$\mathbf{.974 \pm .004}$	$.853 \pm .004$	$.970 \pm .006$
	0.25	$.942 \pm .004$	$.862 \pm .010$	$.941 \pm .004$	$\mathbf{.944 \pm .009}$	$.856 \pm .008$	$.940 \pm .006$
	0.30	$\mathbf{.913 \pm .007}$	$.858 \pm .008$	$.906 \pm .009$	$.903 \pm .010$	$.860 \pm .007$	$.909 \pm .007$
	0.35	$.874 \pm .003$	$.862 \pm .008$	$\mathbf{.878 \pm .007}$	$.878 \pm .002$	$.851 \pm .012$	$.871 \pm .005$
	0.40	$.860 \pm .010$	$.863 \pm .009$	$\mathbf{.863 \pm .010}$	$.858 \pm .009$	$.859 \pm .011$	$.854 \pm .009$
	0.45	$.856 \pm .009$	$\mathbf{.865 \pm .007}$	$.852 \pm .007$	$.855 \pm .007$	$.858 \pm .009$	$.859 \pm .008$
	0.50	$.857 \pm .004$	$.851 \pm .004$	$.854 \pm .004$	$.855 \pm .009$	$.858 \pm .007$	$\mathbf{.861 \pm .006}$
$CovTask$	0.00	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.01	0.000 ± 0.000	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.007 \pm .016$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.119 \pm .095$	$.292 \pm .264$	$\mathbf{1.000 \pm 0.000}$	$.304 \pm .181$	$.383 \pm .114$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.610 \pm .200$	$.741 \pm .180$	$\mathbf{1.000 \pm 0.000}$	$.629 \pm .178$	$.661 \pm .164$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.716 \pm .150$	$.886 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.826 \pm .054$	$.908 \pm .031$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.910 \pm .077$	$.950 \pm .018$	$\mathbf{1.000 \pm 0.000}$	$.910 \pm .053$	$.965 \pm .026$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.993 \pm .003$	$.987 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.952 \pm .030$	$.974 \pm .042$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.988 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.988 \pm .018$	$.992 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.999 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.40	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.45	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.001 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.001 \pm .001$	$.001 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.001 \pm .002$
	0.01	$.017 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.021 \pm .009$	$.017 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.019 \pm .008$
	0.05	$.146 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.113 \pm .025$	$.143 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.122 \pm .019$
	0.10	$.264 \pm .045$	$\mathbf{1.000 \pm 0.000}$	$.259 \pm .027$	$.271 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.273 \pm .031$
	0.15	$.422 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.398 \pm .049$	$.420 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.395 \pm .043$
	0.20	$.528 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.527 \pm .038$	$.535 \pm .029$	$\mathbf{1.000 \pm 0.000}$	$.535 \pm .020$
	0.25	$.715 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.692 \pm .026$	$.683 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.687 \pm .019$
	0.30	$.815 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.821 \pm .021$	$.843 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.825 \pm .027$
	0.35	$.946 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.940 \pm .013$	$.938 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.924 \pm .030$
	0.40	$.982 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.984 \pm .006$	$.984 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.988 \pm .009$
	0.45	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .003$	$.997 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .001$
	0.50	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$

Table 27: Results for the `completeness` dataset when not allowing for shared parameters with independent training using CE, and considering oracle task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
AccTask	0.00	.988 ± .008	.933 ± .020	.893 ± .033	**.990 ± .005**	.943 ± .014	.911 ± .007
	0.01	.987 ± .006	.926 ± .019	.899 ± .020	**.994 ± .004**	.934 ± .026	.908 ± .021
	0.05	**.989 ± .007**	.917 ± .009	.899 ± .024	.981 ± .008	.921 ± .014	.894 ± .027
	0.10	.975 ± .008	.897 ± .015	.905 ± .020	**.982 ± .006**	.916 ± .023	.887 ± .028
	0.15	.976 ± .007	.918 ± .018	.905 ± .018	**.979 ± .007**	.891 ± .015	.896 ± .004
	0.20	.948 ± .028	.889 ± .020	.907 ± .008	**.958 ± .022**	.905 ± .018	.909 ± .009
	0.25	.960 ± .010	.880 ± .007	.886 ± .034	**.961 ± .007**	.886 ± .010	.905 ± .008
	0.30	**.955 ± .012**	.893 ± .014	.887 ± .018	.953 ± .014	.877 ± .020	.884 ± .014
	0.35	**.952 ± .009**	.875 ± .008	.876 ± .020	.944 ± .005	.876 ± .018	.886 ± .015
	0.40	.920 ± .017	.876 ± .011	.879 ± .018	**.926 ± .019**	.865 ± .022	.879 ± .013
	0.45	**.927 ± .009**	.867 ± .012	.883 ± .020	.916 ± .022	.873 ± .016	.873 ± .017
	0.50	.900 ± .009	.868 ± .006	.865 ± .010	**.904 ± .010**	.864 ± .021	.861 ± .015
AccConc	0.00	**.992 ± .002**	.845 ± .008	.991 ± .004	.991 ± .002	.842 ± .008	.990 ± .002
	0.01	**.990 ± .002**	.843 ± .010	.988 ± .003	.989 ± .003	.843 ± .005	.990 ± .002
	0.05	**.987 ± .003**	.845 ± .007	.987 ± .002	.985 ± .005	.844 ± .003	.986 ± .004
	0.10	.980 ± .007	.843 ± .010	.980 ± .006	.980 ± .007	.846 ± .006	**.986 ± .004**
	0.15	.975 ± .005	.847 ± .008	**.977 ± .007**	.974 ± .006	.847 ± .004	.972 ± .003
	0.20	.966 ± .003	.844 ± .006	.968 ± .004	**.970 ± .006**	.848 ± .007	.966 ± .002
	0.25	.959 ± .007	.845 ± .003	**.969 ± .004**	.961 ± .005	.845 ± .007	.967 ± .007
	0.30	**.960 ± .009**	.849 ± .008	.955 ± .006	.956 ± .005	.841 ± .007	.955 ± .006
	0.35	.949 ± .006	.845 ± .004	.946 ± .002	**.953 ± .005**	.844 ± .012	.944 ± .008
	0.40	.937 ± .003	.842 ± .016	.937 ± .003	.941 ± .010	.847 ± .007	**.942 ± .011**
	0.45	.929 ± .008	.845 ± .012	.928 ± .007	**.930 ± .007**	.844 ± .012	.922 ± .007
	0.50	**.912 ± .008**	.851 ± .007	.910 ± .004	.907 ± .007	.841 ± .008	.910 ± .006
CovTask	0.00	.472 ± .077	.461 ± .082	**1.000 ± 0.000**	.410 ± .024	.509 ± .059	**1.000 ± 0.000**
	0.01	.500 ± .064	.569 ± .158	**1.000 ± 0.000**	.485 ± .041	.499 ± .092	**1.000 ± 0.000**
	0.05	.541 ± .038	.684 ± .048	**1.000 ± 0.000**	.551 ± .083	.631 ± .046	**1.000 ± 0.000**
	0.10	.635 ± .042	.742 ± .027	**1.000 ± 0.000**	.614 ± .041	.689 ± .053	**1.000 ± 0.000**
	0.15	.675 ± .079	.738 ± .069	**1.000 ± 0.000**	.589 ± .116	.791 ± .021	**1.000 ± 0.000**
	0.20	.803 ± .059	.840 ± .035	**1.000 ± 0.000**	.772 ± .045	.789 ± .049	**1.000 ± 0.000**
	0.25	.793 ± .037	.864 ± .020	**1.000 ± 0.000**	.803 ± .017	.845 ± .034	**1.000 ± 0.000**
	0.30	.827 ± .027	.854 ± .031	**1.000 ± 0.000**	.809 ± .054	.865 ± .041	**1.000 ± 0.000**
	0.35	.843 ± .012	.892 ± .014	**1.000 ± 0.000**	.854 ± .023	.902 ± .028	**1.000 ± 0.000**
	0.40	.878 ± .033	.884 ± .049	**1.000 ± 0.000**	.867 ± .026	.898 ± .031	**1.000 ± 0.000**
	0.45	.871 ± .024	.905 ± .027	**1.000 ± 0.000**	.898 ± .076	.907 ± .020	**1.000 ± 0.000**
	0.50	.920 ± .020	.922 ± .009	**1.000 ± 0.000**	.881 ± .022	.928 ± .018	**1.000 ± 0.000**
CovConc	0.00	.255 ± .012	**1.000 ± 0.000**	.254 ± .019	.259 ± .013	**1.000 ± 0.000**	.276 ± .012
	0.01	.274 ± .012	**1.000 ± 0.000**	.283 ± .006	.274 ± .007	**1.000 ± 0.000**	.279 ± .006
	0.05	.309 ± .016	**1.000 ± 0.000**	.311 ± .018	.306 ± .013	**1.000 ± 0.000**	.310 ± .017
	0.10	.375 ± .021	**1.000 ± 0.000**	.368 ± .019	.362 ± .011	**1.000 ± 0.000**	.371 ± .012
	0.15	.439 ± .026	**1.000 ± 0.000**	.416 ± .014	.438 ± .017	**1.000 ± 0.000**	.429 ± .012
	0.20	.469 ± .009	**1.000 ± 0.000**	.477 ± .022	.492 ± .025	**1.000 ± 0.000**	.491 ± .014
	0.25	.534 ± .015	**1.000 ± 0.000**	.544 ± .014	.532 ± .017	**1.000 ± 0.000**	.524 ± .018
	0.30	.582 ± .016	**1.000 ± 0.000**	.592 ± .013	.580 ± .014	**1.000 ± 0.000**	.596 ± .013
	0.35	.647 ± .017	**1.000 ± 0.000**	.648 ± .020	.648 ± .013	**1.000 ± 0.000**	.655 ± .010
	0.40	.702 ± .009	**1.000 ± 0.000**	.716 ± .020	.698 ± .009	**1.000 ± 0.000**	.697 ± .013
	0.45	.764 ± .014	**1.000 ± 0.000**	.762 ± .015	.760 ± .007	**1.000 ± 0.000**	.764 ± .014
	0.50	.828 ± .013	**1.000 ± 0.000**	.825 ± .006	.832 ± .009	**1.000 ± 0.000**	.825 ± .005

Table 28: Results for the `completeness` dataset when not allowing for shared parameters with joint training using CE, and considering oracle task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
AccTask	0.00	**.989 ± .011**	.941 ± .022	.891 ± .023	.986 ± .007	.945 ± .019	.890 ± .019
	0.01	**.985 ± .015**	.965 ± .013	.906 ± .023	.983 ± .015	.956 ± .019	.879 ± .025
	0.05	**.981 ± .015**	.939 ± .019	.871 ± .026	.980 ± .013	.933 ± .014	.903 ± .008
	0.10	.965 ± .007	.921 ± .011	.869 ± .014	**.977 ± .010**	.915 ± .013	.899 ± .018
	0.15	.960 ± .015	.902 ± .013	.884 ± .013	**.966 ± .016**	.911 ± .019	.899 ± .021
	0.20	.946 ± .024	.895 ± .019	.889 ± .015	**.956 ± .018**	.896 ± .007	.885 ± .026
	0.25	.948 ± .013	.890 ± .009	.888 ± .011	**.960 ± .017**	.887 ± .014	.882 ± .027
	0.30	**.948 ± .007**	.881 ± .020	.883 ± .021	.941 ± .002	.883 ± .017	.887 ± .022
	0.35	**.943 ± .008**	.885 ± .015	.883 ± .027	.929 ± .020	.866 ± .018	.883 ± .010
	0.40	.920 ± .010	.870 ± .015	.880 ± .006	**.935 ± .020**	.871 ± .017	.874 ± .007
	0.45	**.931 ± .016**	.867 ± .021	.880 ± .012	.923 ± .014	.869 ± .012	.870 ± .018
	0.50	**.915 ± .020**	.871 ± .014	.865 ± .013	.910 ± .017	.857 ± .014	.867 ± .016
AccConc	0.00	**.990 ± .002**	.840 ± .009	.988 ± .002	.988 ± .005	.838 ± .006	.990 ± .003
	0.01	.988 ± .007	.839 ± .005	.984 ± .004	**.989 ± .002**	.843 ± .011	.987 ± .004
	0.05	.986 ± .003	.843 ± .008	**.987 ± .002**	.987 ± .005	.844 ± .004	.982 ± .004
	0.10	.977 ± .004	.841 ± .007	**.981 ± .004**	.980 ± .004	.834 ± .006	.978 ± .002
	0.15	**.978 ± .007**	.831 ± .007	.973 ± .005	.972 ± .006	.836 ± .009	.972 ± .004
	0.20	**.969 ± .005**	.841 ± .009	.967 ± .004	.966 ± .005	.844 ± .007	.964 ± .005
	0.25	**.961 ± .003**	.842 ± .006	.958 ± .006	.960 ± .007	.846 ± .004	.958 ± .003
	0.30	.953 ± .007	.844 ± .006	.953 ± .005	.956 ± .002	.838 ± .008	**.956 ± .010**
	0.35	.946 ± .004	.832 ± .008	**.948 ± .004**	.944 ± .006	.838 ± .006	.946 ± .005
	0.40	.934 ± .008	.839 ± .011	.936 ± .003	.938 ± .007	.837 ± .008	**.941 ± .005**
	0.45	.927 ± .006	.837 ± .006	.926 ± .006	**.930 ± .005**	.843 ± .008	.926 ± .001
	0.50	.909 ± .006	.836 ± .012	**.910 ± .010**	.910 ± .007	.838 ± .009	.910 ± .010
CovTask	0.00	.397 ± .149	.487 ± .118	**1.000 ± 0.000**	.493 ± .075	.503 ± .126	**1.000 ± 0.000**
	0.01	.466 ± .109	.424 ± .131	**1.000 ± 0.000**	.519 ± .093	.458 ± .110	**1.000 ± 0.000**
	0.05	.505 ± .078	.568 ± .119	**1.000 ± 0.000**	.552 ± .083	.580 ± .093	**1.000 ± 0.000**
	0.10	.704 ± .046	.619 ± .050	**1.000 ± 0.000**	.616 ± .034	.683 ± .064	**1.000 ± 0.000**
	0.15	.732 ± .045	.759 ± .053	**1.000 ± 0.000**	.736 ± .053	.707 ± .103	**1.000 ± 0.000**
	0.20	.790 ± .040	.833 ± .028	**1.000 ± 0.000**	.804 ± .041	.834 ± .047	**1.000 ± 0.000**
	0.25	.832 ± .015	.853 ± .008	**1.000 ± 0.000**	.785 ± .054	.844 ± .038	**1.000 ± 0.000**
	0.30	.834 ± .034	.867 ± .029	**1.000 ± 0.000**	.826 ± .041	.868 ± .022	**1.000 ± 0.000**
	0.35	.838 ± .021	.864 ± .032	**1.000 ± 0.000**	.866 ± .039	.917 ± .019	**1.000 ± 0.000**
	0.40	.877 ± .026	.917 ± .035	**1.000 ± 0.000**	.854 ± .024	.905 ± .023	**1.000 ± 0.000**
	0.45	.859 ± .016	.907 ± .033	**1.000 ± 0.000**	.876 ± .022	.912 ± .043	**1.000 ± 0.000**
	0.50	.913 ± .059	.897 ± .007	**1.000 ± 0.000**	.896 ± .030	.930 ± .022	**1.000 ± 0.000**
CovConc	0.00	.257 ± .015	**1.000 ± 0.000**	.261 ± .017	.267 ± .014	**1.000 ± 0.000**	.245 ± .014
	0.01	.266 ± .009	**1.000 ± 0.000**	.274 ± .019	.254 ± .010	**1.000 ± 0.000**	.260 ± .009
	0.05	.301 ± .019	**1.000 ± 0.000**	.310 ± .024	.301 ± .019	**1.000 ± 0.000**	.309 ± .012
	0.10	.363 ± .014	**1.000 ± 0.000**	.340 ± .019	.357 ± .014	**1.000 ± 0.000**	.359 ± .011
	0.15	.396 ± .018	**1.000 ± 0.000**	.409 ± .023	.411 ± .028	**1.000 ± 0.000**	.412 ± .015
	0.20	.466 ± .010	**1.000 ± 0.000**	.472 ± .024	.462 ± .031	**1.000 ± 0.000**	.464 ± .024
	0.25	.516 ± .027	**1.000 ± 0.000**	.507 ± .021	.528 ± .015	**1.000 ± 0.000**	.522 ± .024
	0.30	.567 ± .020	**1.000 ± 0.000**	.556 ± .012	.563 ± .034	**1.000 ± 0.000**	.574 ± .015
	0.35	.619 ± .007	**1.000 ± 0.000**	.628 ± .005	.623 ± .015	**1.000 ± 0.000**	.618 ± .019
	0.40	.690 ± .026	**1.000 ± 0.000**	.679 ± .011	.676 ± .025	**1.000 ± 0.000**	.674 ± .012
	0.45	.745 ± .016	**1.000 ± 0.000**	.732 ± .006	.742 ± .008	**1.000 ± 0.000**	.741 ± .013
	0.50	.815 ± .007	**1.000 ± 0.000**	.801 ± .011	.795 ± .013	**1.000 ± 0.000**	.806 ± .017

Table 29: Results for the `completeness` dataset when not allowing for shared parameters with independent training using OVA, and considering oracle task expert and oracle concept expert. LS refers to the label-smoothing-free implementation, while NLS to the one with label smoothing. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .007$	$.906 \pm .032$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$.925 \pm .010$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.987 \pm .013$	$.914 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.999 \pm .002$	$.918 \pm .021$
	0.05	$.998 \pm .004$	$.965 \pm .024$	$.914 \pm .020$	$\mathbf{.999 \pm .002}$	$.969 \pm .017$	$.901 \pm .024$
	0.10	$.973 \pm .019$	$.899 \pm .007$	$.914 \pm .020$	$\mathbf{.987 \pm .004}$	$.910 \pm .026$	$.900 \pm .029$
	0.15	$.961 \pm .014$	$.902 \pm .015$	$.907 \pm .024$	$\mathbf{.969 \pm .018}$	$.880 \pm .017$	$.907 \pm .009$
	0.20	$\mathbf{.931 \pm .016}$	$.871 \pm .011$	$.893 \pm .009$	$.926 \pm .023$	$.883 \pm .010$	$.896 \pm .007$
	0.25	$.908 \pm .009$	$.866 \pm .010$	$.875 \pm .026$	$\mathbf{.915 \pm .012}$	$.867 \pm .017$	$.887 \pm .010$
	0.30	$\mathbf{.892 \pm .018}$	$.863 \pm .016$	$.875 \pm .014$	$.889 \pm .010$	$.860 \pm .013$	$.867 \pm .011$
	0.35	$.885 \pm .020$	$.849 \pm .014$	$.862 \pm .009$	$\mathbf{.897 \pm .008}$	$.855 \pm .008$	$.864 \pm .018$
	0.40	$.860 \pm .017$	$.846 \pm .018$	$.848 \pm .015$	$\mathbf{.874 \pm .012}$	$.846 \pm .016$	$.860 \pm .013$
	0.45	$\mathbf{.859 \pm .013}$	$.836 \pm .014$	$.844 \pm .010$	$.847 \pm .009$	$.840 \pm .014$	$.844 \pm .016$
	0.50	$.830 \pm .014$	$.832 \pm .008$	$\mathbf{.834 \pm .016}$	$.813 \pm .014$	$.832 \pm .006$	$.824 \pm .010$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.845 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.842 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.843 \pm .010$	1.000 ± 0.000	1.000 ± 0.000	$.843 \pm .005$	1.000 ± 0.000
	0.05	$.998 \pm .001$	$.845 \pm .007$	$\mathbf{.999 \pm .001}$	$.998 \pm 0.000$	$.844 \pm .003$	$.997 \pm .001$
	0.10	$.993 \pm .001$	$.843 \pm .010$	$.994 \pm 0.000$	$.993 \pm .002$	$.846 \pm .006$	$\mathbf{.994 \pm .002}$
	0.15	$\mathbf{.987 \pm .003}$	$.847 \pm .008$	$.984 \pm .003$	$.983 \pm .002$	$.847 \pm .004$	$.984 \pm .003$
	0.20	$.969 \pm .007$	$.844 \pm .006$	$\mathbf{.969 \pm .004}$	$.967 \pm .005$	$.848 \pm .007$	$.969 \pm .004$
	0.25	$.950 \pm .004$	$.845 \pm .003$	$.949 \pm .006$	$.949 \pm .003$	$.845 \pm .007$	$\mathbf{.954 \pm .009}$
	0.30	$\mathbf{.937 \pm .009}$	$.849 \pm .008$	$.931 \pm .006$	$.935 \pm .005$	$.841 \pm .007$	$.934 \pm .005$
	0.35	$.913 \pm .003$	$.845 \pm .004$	$.914 \pm .005$	$\mathbf{.917 \pm .005}$	$.844 \pm .012$	$.911 \pm .005$
	0.40	$.891 \pm .004$	$.842 \pm .016$	$.891 \pm .005$	$.891 \pm .007$	$.847 \pm .007$	$\mathbf{.892 \pm .007}$
	0.45	$\mathbf{.868 \pm .008}$	$.845 \pm .012$	$.866 \pm .007$	$.864 \pm .007$	$.844 \pm .012$	$.863 \pm .005$
	0.50	$.837 \pm .005$	$\mathbf{.851 \pm .007}$	$.837 \pm .007$	$.836 \pm .004$	$.841 \pm .008$	$.839 \pm .005$
$CovTask$	0.00	0.000 ± 0.000	$.026 \pm .058$	$\mathbf{1.000 \pm 0.000}$	0.000 ± 0.000	$.009 \pm .015$	$\mathbf{1.000 \pm 0.000}$
	0.01	$.077 \pm .070$	$.110 \pm .092$	$\mathbf{1.000 \pm 0.000}$	$.071 \pm .081$	$.019 \pm .042$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.335 \pm .241$	$.449 \pm .171$	$\mathbf{1.000 \pm 0.000}$	$.300 \pm .046$	$.407 \pm .127$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.668 \pm .011$	$.753 \pm .072$	$\mathbf{1.000 \pm 0.000}$	$.608 \pm .076$	$.773 \pm .070$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.728 \pm .012$	$.819 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.680 \pm .137$	$.872 \pm .035$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.839 \pm .064$	$.896 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.832 \pm .060$	$.883 \pm .040$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.897 \pm .024$	$.910 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$.896 \pm .019$	$.899 \pm .035$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.910 \pm .012$	$.924 \pm .024$	$\mathbf{1.000 \pm 0.000}$	$.918 \pm .012$	$.919 \pm .019$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.939 \pm .029$	$.961 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.929 \pm .012$	$.947 \pm .020$	$\mathbf{1.000 \pm 0.000}$
	0.40	$.957 \pm .034$	$.960 \pm .029$	$\mathbf{1.000 \pm 0.000}$	$.951 \pm .031$	$.952 \pm .014$	$\mathbf{1.000 \pm 0.000}$
	0.45	$.970 \pm .020$	$.987 \pm .010$	$\mathbf{1.000 \pm 0.000}$	$.990 \pm .009$	$.979 \pm .021$	$\mathbf{1.000 \pm 0.000}$
	0.50	$.997 \pm .004$	$.997 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.996 \pm .007$	$.997 \pm .004$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.002 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.004 \pm .002$	$.008 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.005 \pm .004$
	0.01	$.026 \pm .005$	$\mathbf{1.000 \pm 0.000}$	$.029 \pm .016$	$.037 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.024 \pm .008$
	0.05	$.117 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.129 \pm .043$	$.138 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.144 \pm .017$
	0.10	$.274 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.257 \pm .026$	$.250 \pm .005$	$\mathbf{1.000 \pm 0.000}$	$.253 \pm .034$
	0.15	$.347 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.357 \pm .036$	$.384 \pm .016$	$\mathbf{1.000 \pm 0.000}$	$.363 \pm .013$
	0.20	$.470 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.481 \pm .027$	$.492 \pm .034$	$\mathbf{1.000 \pm 0.000}$	$.472 \pm .034$
	0.25	$.603 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.605 \pm .034$	$.598 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.603 \pm .030$
	0.30	$.677 \pm .035$	$\mathbf{1.000 \pm 0.000}$	$.711 \pm .016$	$.693 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.695 \pm .025$
	0.35	$.789 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.778 \pm .015$	$.782 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.790 \pm .009$
	0.40	$.862 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.864 \pm .009$	$.862 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.853 \pm .007$
	0.45	$.930 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.932 \pm .007$	$.930 \pm .005$	$\mathbf{1.000 \pm 0.000}$	$.927 \pm .007$
	0.50	$.986 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.990 \pm .002$	$.989 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.989 \pm .002$

Table 30: Results for the `completeness` dataset when not allowing for shared parameters with joint training using OVA, and considering oracle task expert and oracle concept expert. We report $avg \pm std$ and highlight the best baseline in bold.

Metric	λ	DCBM-LS	DCBM-NC-LS	DCBM-NT-LS	DCBM-NLS	DCBM-NC-NLS	DCBM-NT-NLS
$AccTask$	0.00	$\mathbf{1.000 \pm 0.000}$	$.997 \pm .004$	$.907 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .004$	$.927 \pm .022$
	0.01	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.933 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.913 \pm .024$
	0.05	$\mathbf{.998 \pm .004}$	$.984 \pm .010$	$.903 \pm .022$	$.996 \pm .004$	$.968 \pm .021$	$.923 \pm .006$
	0.10	$\mathbf{.980 \pm .005}$	$.925 \pm .013$	$.907 \pm .031$	$.970 \pm .029$	$.915 \pm .020$	$.918 \pm .009$
	0.15	$\mathbf{.958 \pm .015}$	$.879 \pm .017$	$.899 \pm .016$	$.956 \pm .008$	$.877 \pm .017$	$.910 \pm .009$
	0.20	$.927 \pm .018$	$.879 \pm .018$	$.893 \pm .022$	$\mathbf{.942 \pm .008}$	$.883 \pm .015$	$.897 \pm .020$
	0.25	$.926 \pm .016$	$.870 \pm .011$	$.893 \pm .015$	$\mathbf{.928 \pm .012}$	$.861 \pm .011$	$.894 \pm .016$
	0.30	$\mathbf{.928 \pm .010}$	$.862 \pm .016$	$.880 \pm .011$	$.922 \pm .024$	$.849 \pm .016$	$.882 \pm .015$
	0.35	$\mathbf{.916 \pm .010}$	$.869 \pm .019$	$.876 \pm .019$	$.909 \pm .019$	$.843 \pm .012$	$.877 \pm .014$
	0.40	$.888 \pm .014$	$.847 \pm .017$	$.874 \pm .017$	$\mathbf{.895 \pm .018}$	$.848 \pm .012$	$.873 \pm .018$
	0.45	$.871 \pm .016$	$.832 \pm .008$	$.881 \pm .017$	$\mathbf{.883 \pm .009}$	$.841 \pm .009$	$.859 \pm .004$
	0.50	$.861 \pm .016$	$.825 \pm .009$	$.851 \pm .017$	$\mathbf{.865 \pm .016}$	$.829 \pm .008$	$.862 \pm .015$
$AccConc$	0.00	$\mathbf{1.000 \pm 0.000}$	$.836 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$\mathbf{1.000 \pm 0.000}$	$.832 \pm .006$	$\mathbf{1.000 \pm 0.000}$
	0.01	$\mathbf{1.000 \pm 0.000}$	$.835 \pm .008$	$\mathbf{1.000 \pm 0.000}$	1.000 ± 0.000	$.840 \pm .013$	$\mathbf{1.000 \pm 0.000}$
	0.05	$.998 \pm .001$	$.840 \pm .008$	$.998 \pm .001$	$\mathbf{1.000 \pm 0.000}$	$.839 \pm .003$	$.998 \pm .001$
	0.10	$.991 \pm .004$	$.838 \pm .004$	$\mathbf{.995 \pm .002}$	$.994 \pm .001$	$.830 \pm .006$	$.994 \pm .001$
	0.15	$.982 \pm .006$	$.826 \pm .008$	$\mathbf{.985 \pm .004}$	$.984 \pm .003$	$.829 \pm .008$	$.981 \pm .005$
	0.20	$.964 \pm .003$	$.837 \pm .008$	$.967 \pm .003$	$.964 \pm .007$	$.839 \pm .008$	$\mathbf{.969 \pm .005}$
	0.25	$.947 \pm .003$	$.836 \pm .005$	$.949 \pm .008$	$.950 \pm .008$	$.840 \pm .006$	$\mathbf{.952 \pm .002}$
	0.30	$.936 \pm .007$	$.843 \pm .006$	$\mathbf{.938 \pm .006}$	$.937 \pm .005$	$.834 \pm .006$	$.937 \pm .004$
	0.35	$.914 \pm .003$	$.830 \pm .004$	$.918 \pm .005$	$.917 \pm .003$	$.829 \pm .005$	$\mathbf{.920 \pm .006}$
	0.40	$.891 \pm .005$	$.832 \pm .012$	$.896 \pm .005$	$\mathbf{.901 \pm .008}$	$.831 \pm .006$	$.900 \pm .007$
	0.45	$.872 \pm .006$	$.835 \pm .003$	$.873 \pm .003$	$\mathbf{.875 \pm .008}$	$.837 \pm .011$	$.875 \pm .005$
	0.50	$\mathbf{.855 \pm .004}$	$.830 \pm .008$	$.853 \pm .005$	$.855 \pm .008$	$.834 \pm .004$	$.849 \pm .008$
$CovTask$	0.00	0.000 ± 0.000	$.014 \pm .013$	$\mathbf{1.000 \pm 0.000}$	$.048 \pm .046$	$.010 \pm .022$	$\mathbf{1.000 \pm 0.000}$
	0.01	0.000 ± 0.000	$.001 \pm .002$	$\mathbf{1.000 \pm 0.000}$	$.019 \pm .042$	0.000 ± 0.000	$\mathbf{1.000 \pm 0.000}$
	0.05	$.317 \pm .076$	$.223 \pm .095$	$\mathbf{1.000 \pm 0.000}$	$.249 \pm .152$	$.360 \pm .169$	$\mathbf{1.000 \pm 0.000}$
	0.10	$.678 \pm .091$	$.657 \pm .064$	$\mathbf{1.000 \pm 0.000}$	$.683 \pm .069$	$.746 \pm .088$	$\mathbf{1.000 \pm 0.000}$
	0.15	$.783 \pm .032$	$.873 \pm .023$	$\mathbf{1.000 \pm 0.000}$	$.770 \pm .031$	$.861 \pm .050$	$\mathbf{1.000 \pm 0.000}$
	0.20	$.853 \pm .018$	$.885 \pm .038$	$\mathbf{1.000 \pm 0.000}$	$.854 \pm .005$	$.876 \pm .026$	$\mathbf{1.000 \pm 0.000}$
	0.25	$.876 \pm .029$	$.919 \pm .025$	$\mathbf{1.000 \pm 0.000}$	$.875 \pm .029$	$.912 \pm .025$	$\mathbf{1.000 \pm 0.000}$
	0.30	$.894 \pm .019$	$.924 \pm .022$	$\mathbf{1.000 \pm 0.000}$	$.901 \pm .044$	$.945 \pm .016$	$\mathbf{1.000 \pm 0.000}$
	0.35	$.924 \pm .019$	$.921 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.922 \pm .032$	$.969 \pm .012$	$\mathbf{1.000 \pm 0.000}$
	0.40	$.961 \pm .016$	$.968 \pm .020$	$\mathbf{1.000 \pm 0.000}$	$.960 \pm .022$	$.960 \pm .023$	$\mathbf{1.000 \pm 0.000}$
	0.45	$.989 \pm .014$	$.975 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.978 \pm .016$	$.989 \pm .008$	$\mathbf{1.000 \pm 0.000}$
	0.50	$\mathbf{1.000 \pm 0.000}$	$.995 \pm .009$	$\mathbf{1.000 \pm 0.000}$	$.998 \pm .004$	$.996 \pm .005$	$\mathbf{1.000 \pm 0.000}$
$CovConc$	0.00	$.007 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.004 \pm .003$	$.006 \pm .004$	$\mathbf{1.000 \pm 0.000}$	$.003 \pm .002$
	0.01	$.022 \pm .003$	$\mathbf{1.000 \pm 0.000}$	$.027 \pm .010$	$.018 \pm .008$	$\mathbf{1.000 \pm 0.000}$	$.023 \pm .007$
	0.05	$.131 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.134 \pm .026$	$.108 \pm .037$	$\mathbf{1.000 \pm 0.000}$	$.137 \pm .018$
	0.10	$.230 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.229 \pm .017$	$.236 \pm .036$	$\mathbf{1.000 \pm 0.000}$	$.244 \pm .018$
	0.15	$.341 \pm .049$	$\mathbf{1.000 \pm 0.000}$	$.338 \pm .042$	$.364 \pm .039$	$\mathbf{1.000 \pm 0.000}$	$.377 \pm .018$
	0.20	$.480 \pm .011$	$\mathbf{1.000 \pm 0.000}$	$.469 \pm .023$	$.453 \pm .057$	$\mathbf{1.000 \pm 0.000}$	$.457 \pm .021$
	0.25	$.589 \pm .039$	$\mathbf{1.000 \pm 0.000}$	$.587 \pm .026$	$.568 \pm .026$	$\mathbf{1.000 \pm 0.000}$	$.575 \pm .020$
	0.30	$.650 \pm .028$	$\mathbf{1.000 \pm 0.000}$	$.661 \pm .026$	$.648 \pm .033$	$\mathbf{1.000 \pm 0.000}$	$.680 \pm .023$
	0.35	$.741 \pm .015$	$\mathbf{1.000 \pm 0.000}$	$.761 \pm .014$	$.745 \pm .021$	$\mathbf{1.000 \pm 0.000}$	$.764 \pm .013$
	0.40	$.813 \pm .019$	$\mathbf{1.000 \pm 0.000}$	$.829 \pm .009$	$.791 \pm .048$	$\mathbf{1.000 \pm 0.000}$	$.820 \pm .025$
	0.45	$.882 \pm .014$	$\mathbf{1.000 \pm 0.000}$	$.896 \pm .013$	$.882 \pm .006$	$\mathbf{1.000 \pm 0.000}$	$.893 \pm .015$
	0.50	$.932 \pm .007$	$\mathbf{1.000 \pm 0.000}$	$.951 \pm .012$	$.931 \pm .017$	$\mathbf{1.000 \pm 0.000}$	$.953 \pm .011$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification:We present Deferring Concept Bottleneck Models, as a method to bridge Learning to Defer (L2D) and Concept Bottleneck Models (CBMs). We provide theoretical results and we empirically show the benefits of our approach in addressing open issues in CBMs and L2D.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Section 6, we discuss how treating concepts to be independent without considering interactions among them might not be realistic. While this limitation is shared with standard CBMs, it also affects DCBMs.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Yes, all the proofs are reported in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All experimental details are reported in Appendix B and documented in the code attached to the submission.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code is available at `https://github.com/andrepugni/DCBM`

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Yes, the experimental details are reported in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report the number of repetitions and plot standard deviation for our methods.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report such details in Appendix B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We test our methods on synthetic data or widely known datasets.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: In Section 6, we discuss how having interpretable models might help audit deferring systems.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release novel datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the authors of used assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We did not perform crowdsourcing.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: No human participants are involved in our study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [No]

Justification: No, the introduced method does not interact with LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.