A Appendix

A.1 Dataset preparation

We split the labels in the label space as seen labels and unseen labels. Unseen labels do not necessarily need to be leaf labels, and if an intermediate label is chosen as unseen, then all its descendant labels will be set as unseen. Meanwhile, each data instance in dev/test sets will contain at least one unseen label.

Table 5 shows the example instances of Yelp, WOS and QCD datasets used in this work.

A.2 Rollback Results with DistilBERT

As shown in Table 6, DistilBERT+RLHR with Rollback algorithm can achieve the best performance on most evaluation metrics. Although the hierarchical inference method can improve DistilBERT on QCD dataset, its performance is not consistent. It lowers the performance by large margins on WOS with both DistilBERT and DistilBERT+RLHR. In contrast, the rollback algorithm has consistent performance on all the three datasets, especially when combined with our proposed RLHR approach.

A.3 Influence of λ with DistilBERT

As shown in Figure 4, the influence of parameter λ on three datasets with DistilBERT is similar to that with BERT. For Yelp and QCD datasets, a larger λ helps achieve better classification performance on unseen labels, while it will bring more logical errors. On the contrary, a relatively small λ yields both better classification performance and lower logical error rates on WOS dataset, as shown in Figure 4b. The results support our analyses in Section 5.3.4.

A.4 Deduction Path Analysis

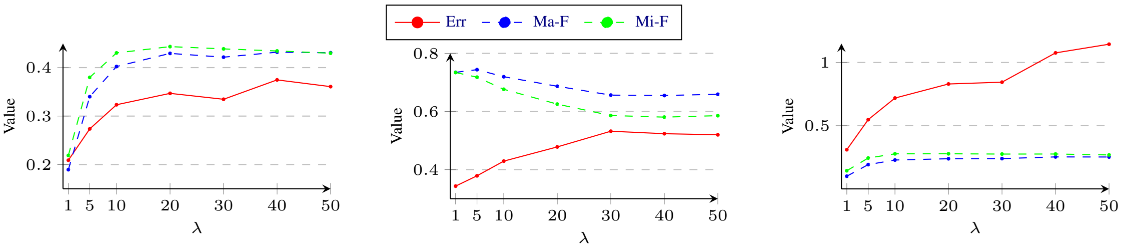
We represent the results of deduction paths in this section, which is an important evaluation of if the model captures the interdependencies of labels. A path is considered as correct when it equals to or belongs to a golden deduction path, and we report Example-based Precision, Recall and F1 based on BERT. As shown in Table 4, BERT can achieve high recall but low precision on the deduction paths, which means that it tends to predict more labels as correct. This is because pretrained models only take the literal tokens of labels as input without any label structure information. On the contrary, RLHR, which incorporates the label hierarchy, can provide more accurate predictions of deduction paths with much higher precision on all the three datasets.

Dataset	BERT			BERT+RLHR		
	P	R	F1	P	R	F1
Yelp	17.17	72.54	26.03	38.04	52.61	40.27
WOS	33.25	77.57	44.35	47.34	66.51	53.28
QCD	18.43	58.37	26.68	22.55	57.11	30.71

Table 4: Performance on deduction paths. P, R, F1 denote Example-based Precision, Recall and F1.

Dataset	Text	Labels
Yelp	Mini donuts at it's finest. I was there on Saturday and it was absolutely delicious. I had a mini six pack of D O's. I would highly recommend this place for a sweet snack. Five thumbs up.	*Food*, *Restaurants*, *Donuts*, *Food Stands*
WOS	This paper presents the design and experimental evaluation of discrete time sliding mode controller using multirate output feedback to minimize structural vibration of a cantilever beam using shape memory alloy wires as control actuators and piezoceramics as sensor and disturbance actuator. Linear dynamic models of the smart cantilever beam are obtained using online recursive least square parameter estimation. A digital control system that consists of Simulink (TM) modeling software and dSPACE DS1104 controller board is used for identification and control. The effectiveness of the controller is shown through simulation and experimentation by exciting the structure at resonance.	*ECE*, *Digital control*
QCD	ipad usb c hub	*Electronics*, *Accessories & Supplies*, *Audio & Video Accessories*

Table 5: Examples of the three datasets



(a) Yelp

(b) WOS

(c) QCD

Figure 4: Influence of λ on RLHR approach with DistilBERT. Err, Ma-F and Mi-F denote logical error rate, Macro-F1 and Micro-F1 respectively.

Method	Setting	Yelp			WOS			QCD		
		Ma-F	Mi-F	EBF	Ma-F	Mi-F	EBF	Ma-F	Mi-F	EBF
DistilBERT	ZS	41.42	40.33	30.44	70.69	65.19	55.18	23.68	24.95	33.57
	GZS	21.29	28.18		68.03	63.64		24.43	34.29	
DistilBERT +Hie-Infe	ZS	41.88	41.00	30.61	67.81	66.45	53.13	21.13	29.29	34.35
	GZS	21.49	28.36		65.65	64.05		23.91	35.12	
DistilBERT +Rollback	ZS	41.49	40.32	30.47	70.69	65.19	56.54	23.81	24.7	33.34
	GZS	21.28	28.18		68.44	63.31		24.36	33.99	
DistilBERT+RLHR	ZS	42.16	43.87	40.85	74.56	72.44	61.06	24.58	27.79	37.46
	GZS	26.95	40.43		71.65	68.05		26.10	38.73	
DistilBERT+RLHR +Hie-Infe	ZS	39.48	41.65	40.65	63.61	64.21	53.39	20.18	**29.68**	**38.13**
	GZS	26.79	40.44		62.63	64.05		24.98	**39.44**	
DistilBERT+RLHR +Rollback	ZS	**42.27**	**43.91**	**41.03**	**74.56**	**72.44**	**65.64**	**24.89**	28.34	37.45
	GZS	**26.97**	**40.55**		**73.14**	**71.48**		**26.17**	38.68	

Table 6: Results and comparisons of our matching-score-based rollback algorithm on DistilBERT. Ma-F, Mi-F, EBF, Err denote Macro-F1, Micro-F1, Example-based F1 and logical error rate respectively, and ZS, GZS denote zero-shot setting and generalized zero-shot setting. Bold figures indicate the best results for each metric.