A Appendix

This document contains only the appendix.

Case	3	3.3	Total	% (3/Total)
A	258	3,266	3,524	7.3
B	332	1,648	1,980	16.8
C	282	1,362	1,644	17.2
D	100	1,012	1,112	9.0
Total	972	7,288	8,260	11.8

Table 1: Number of messages generated with each version of Llama used, broken down by case

Table 2: Advantages and Disadvantages of Each Scenario in Cyberbullying Detection

Scenario	Advantages	Disadvantages
Baseline (Gold-Standard Only)	High-quality, reliable data	High costs and scalability challenges; Requires significant time and expert annotation effort.
LLM as Classifier	No need for labeled data or training; Quick deployment; Handles nuanced language patterns.	Computationally expensive; May be less accurate than fine-tuned classifiers on domain-specific data.
Synthetic Labels for Unlabeled Data	Utilizes existing unlabeled data; Cost-effective dataset creation.	Label quality depends on LLM performance; May require validation to ensure consistency and accuracy.
Fully Synthetic Data	Enables training when no authentic data is available; Suitable for low-resource domains.	Synthetic data may lack diversity and realism; Risk of overfitting to generated patterns.

		Development Set		**Test Set**		
Size	**Sampling**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**	**Rep.**
100%	up	79.8% ± 1.5	76.3% ± 1.8	80.8% ± 1.5	77.3% ± 1.5	50
20%	none	74.2% ± 2.8	68.1% ± 3.4	73.7% ± 2.7	67.7% ± 2.5	65
50%	none	78.8% ± 2.1	74.0% ± 2.6	79.4% ± 1.7	74.5% ± 2.1	65
80%	none	80.0% ± 1.8	75.6% ± 2.2	80.4% ± 1.5	75.8% ± 1.7	65
100%	none	80.9% ± 1.6	76.8% ± 1.8	81.5% ± 1.2	77.0% ± 1.6	85
200%	none	80.8% ± 1.6	76.6% ± 2.0	81.7% ± 1.2	77.2% ± 1.4	50

Table 3: Development and test set results in scenario 1: training BERT-based classifiers on the training split of the authentic data; at least 45 repetitions with different random seeds; also shown for comparison results for training on samples from 20% to 80%, as well as two copies (200%) of the data; both the training set and the development set are sampled to the given relative size of the authentic data split; "Sampling" refers to the strategy for addressing class imbalance in the training data

Rel.		**Development Set**		**Test Set**		
Size	**Sampling**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**	**Rep.**
100%	up	71.5% ± 5.7	62.5% ± 4.4	71.4% ± 4.3	61.8% ± 3.0	50
100%	none	72.2% ± 3.9	58.6% ± 4.5	72.1% ± 3.2	58.9% ± 3.7	50
120%	none	72.8% ± 3.1	59.6% ± 4.3	72.8% ± 2.7	59.8% ± 4.3	65
140%	none	73.7% ± 3.2	61.2% ± 4.1	73.5% ± 2.3	61.1% ± 4.0	65
160%	none	74.4% ± 3.0	62.8% ± 3.5	74.0% ± 2.2	62.3% ± 3.0	65
180%	none	74.7% ± 2.8	63.3% ± 3.8	74.2% ± 2.2	62.5% ± 3.5	65
200%	none	75.2% ± 2.2	64.0% ± 3.0	74.5% ± 1.9	63.1% ± 2.9	65

Table 4: Development and test set results for **Llama3 with default "not harmfull" label** in scenario 3: training a BERT-based classifier on synthetic data matching 100% to 200% of the size available in scenario 1. "Sampling" refers to the strategy for addressing class imbalance in the training data; at least 45 repetitions with different random seeds;

Rel. Size	Sampling	Development Set Accuracy	Development Set Macro-F1	Test Set Accuracy	Test Set Macro-F1	Rep.
100%	up	72.9% ± 4.8	64.7% ± 4.0	72.7% ± 3.5	64.0% ± 3.1	50
200%	up	75.1% ± 3.3	68.1% ± 2.7	74.9% ± 2.8	67.5% ± 2.3	45
100%	none	73.6% ± 4.1	63.8% ± 3.8	73.4% ± 3.0	63.3% ± 3.3	50
120%	none	74.8% ± 3.6	65.3% ± 3.4	74.5% ± 2.5	64.7% ± 3.0	65
140%	none	75.2% ± 3.4	65.9% ± 3.6	75.0% ± 2.3	65.4% ± 3.0	65
160%	none	76.1% ± 2.6	67.3% ± 3.1	75.3% ± 2.2	66.0% ± 2.8	65
180%	none	76.0% ± 2.7	67.2% ± 2.9	75.6% ± 1.9	66.3% ± 2.4	65
200%	none	76.6% ± 2.5	68.3% ± 2.8	75.8% ± 2.1	67.0% ± 2.4	65

Table 5: Development and test set results for **Llama3 with unlabelled messages removed** in scenario 3: training a BERT-based classifier on synthetic data matching 100% to 200% of the size available in scenario 1. “Sampling” refers to the strategy for addressing class imbalance in the training data; at least 45 repetitions with different random seeds

Rel. Size	Sampling	Development Set Accuracy	Development Set Macro-F1	Test Set Accuracy	Test Set Macro-F1	Rep.
100%	up	69.5% ± 1.8	50.7% ± 3.2	71.7% ± 1.6	53.6% ± 4.8	50
100%	none	69.7% ± 0.8	44.4% ± 3.1	70.2% ± 0.9	45.1% ± 4.1	50
120%	none	69.8% ± 0.7	44.3% ± 2.7	70.3% ± 1.0	44.9% ± 3.9	65
140%	none	69.9% ± 0.8	44.2% ± 3.0	70.2% ± 1.0	44.5% ± 3.7	65
160%	none	69.9% ± 0.8	44.3% ± 3.1	70.3% ± 0.9	44.9% ± 3.7	65
180%	none	70.0% ± 0.7	45.1% ± 3.0	70.4% ± 0.9	45.5% ± 3.9	65
200%	none	70.1% ± 0.7	44.9% ± 3.0	70.4% ± 0.9	45.2% ± 3.7	65

Table 6: Development and test set results for **GPT-4o** in scenario 3: training a BERT-based classifier on synthetic data matching 100% to 200% of the size available in scenario 1. “Sampling” refers to the strategy for addressing class imbalance in the training data; at least 45 repetitions with different random seeds

Rel. Size	Sampling	Development Set Accuracy	Development Set Macro-F1	Test Set Accuracy	Test Set Macro-F1	Rep.
100%	up	51.3% ± 8.2	50.7% ± 8.1	53.5% ± 7.8	52.9% ± 7.5	50
100%	none	49.9% ± 8.5	49.2% ± 8.5	52.2% ± 8.0	51.6% ± 7.7	50
120%	none	50.7% ± 8.6	50.0% ± 8.7	52.9% ± 8.1	52.4% ± 7.9	65
140%	none	50.7% ± 8.9	50.0% ± 9.0	52.9% ± 8.1	52.4% ± 7.9	65
160%	none	51.2% ± 8.9	50.6% ± 9.0	53.5% ± 8.1	52.9% ± 7.9	65
180%	none	51.4% ± 8.8	50.8% ± 8.9	53.7% ± 8.3	53.2% ± 8.1	65
200%	none	52.3% ± 8.3	51.8% ± 8.3	54.2% ± 7.8	53.7% ± 7.5	65

Table 7: Development and test set results for **Grok** in scenario 3: training a BERT-based classifier on synthetic data matching 100% to 200% of the size available in scenario 1. “Sampling” refers to the strategy for addressing class imbalance in the training data; at least 45 repetitions with different random seeds

			Development Set		Test Set		
Labels	**Size**	**Sampling**	**Accuracy**	**Macro-F1**	**Accuracy**	**Macro-F1**	**Rep.**
D0	20%	up	72.3% ± 4.3	59.2% ± 12.0	71.0% ± 3.8	58.0% ± 11.2	50
D0	100%	up	78.7% ± 1.1	74.0% ± 1.4	77.9% ± 0.9	72.6% ± 1.3	50
D0	20%	none	74.9% ± 1.9	64.0% ± 5.0	73.1% ± 2.0	61.8% ± 4.8	85
D0	50%	none	78.3% ± 1.7	71.2% ± 2.6	76.4% ± 1.7	68.3% ± 2.5	85
D0	80%	none	79.7% ± 1.4	73.7% ± 1.9	77.3% ± 0.9	70.0% ± 1.2	85
D0	100%	none	80.0% ± 1.2	73.9% ± 1.6	77.6% ± 0.9	70.3% ± 1.3	85
D0	200%	none	80.0% ± 1.2	74.2% ± 1.6	77.5% ± 0.9	70.4% ± 1.4	50
FU	20%	up	73.0% ± 3.5	61.0% ± 12.4	72.3% ± 3.0	60.5% ± 12.1	50
FU	100%	up	79.6% ± 1.0	75.3% ± 1.1	79.2% ± 0.7	74.3% ± 0.9	50
FU	20%	none	74.8% ± 1.9	66.9% ± 2.9	73.7% ± 1.7	65.7% ± 2.5	85
FU	50%	none	79.3% ± 1.5	74.2% ± 1.9	78.0% ± 1.3	72.1% ± 1.7	85
FU	80%	none	80.1% ± 0.9	75.5% ± 1.0	78.8% ± 1.0	73.2% ± 1.2	85
FU	100%	none	80.4% ± 1.0	75.9% ± 1.2	79.1% ± 1.0	73.5% ± 1.2	85
FU	200%	none	80.2% ± 1.1	75.7% ± 1.2	78.8% ± 0.8	73.4% ± 1.0	50
CH	20%	up	71.9% ± 3.0	57.1% ± 10.6	72.5% ± 3.2	58.6% ± 11.6	50
CH	100%	up	77.5% ± 0.9	71.3% ± 1.3	78.3% ± 0.8	72.5% ± 1.1	50
CH	20%	none	72.0% ± 1.7	55.3% ± 6.7	72.6% ± 1.6	56.9% ± 7.1	85
CH	50%	none	75.1% ± 1.7	64.4% ± 3.7	76.4% ± 1.3	66.9% ± 2.7	85
CH	80%	none	76.2% ± 1.1	67.2% ± 2.0	77.2% ± 0.9	69.0% ± 1.7	85
CH	100%	none	77.1% ± 0.9	68.7% ± 1.5	77.9% ± 0.9	70.2% ± 1.4	85
CH	200%	none	77.1% ± 1.0	68.6% ± 2.1	78.0% ± 0.7	70.2% ± 1.3	50
GR	20%	up	70.5% ± 3.4	67.3% ± 3.2	69.4% ± 3.3	66.6% ± 2.6	50
GR	100%	up	75.6% ± 1.4	73.0% ± 1.4	74.6% ± 1.8	71.7% ± 1.4	50
GR	20%	none	71.3% ± 2.6	67.3% ± 2.5	70.4% ± 3.0	66.9% ± 2.2	85
GR	50%	none	75.0% ± 1.6	72.0% ± 1.7	74.2% ± 2.0	71.0% ± 1.7	85
GR	80%	none	76.1% ± 1.1	73.3% ± 1.2	74.4% ± 1.6	71.3% ± 1.3	85
GR	100%	none	76.3% ± 1.2	73.7% ± 1.2	75.2% ± 1.7	72.1% ± 1.4	85
GR	200%	none	76.3% ± 1.2	73.6% ± 1.2	75.1% ± 1.5	72.1% ± 1.3	50

Table 8: Development and test set results in scenario 4: training BERT-based classifiers on the training split of the authentic data with synthetic labels predicted by (a) D0 = Llama3, assuming “Not Harmful” when no label is found in the LLM output, (b) FU = Llama3, removing dataset items for which no label is found in the LLM output, (c) CH = ChatGPT and (d) GR = Grok; at least 45 repetitions with different random seeds; also shown for comparison results for training on samples from 20% to 80%, as well as tewo copies (200%) of the data; both the training set and the development set are sampled to the given relative size of the authentic data split; “Sampling” refers to the strategy for addressing class imbalance in the training data