LG-CoTrain — Results Dashboard

Semi-supervised co-training · BERT · HumAID dataset · 2 result sets · 147 total experiments

Data Analysis gpt-4o-run-1 gpt-4o-run-2

California Wildfires 2018

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	97	14	28	5	10	25	50	92	87	72	47
displaced_people_and_evacuations	258	38	72	5	10	25	50	253	248	233	208
infrastructure_and_utility_damage	295	43	84	5	10	25	50	290	285	270	245
injured_or_dead_people	1362	199	385	5	10	25	50	1357	1352	1337	1312
missing_or_found_people	125	18	36	5	10	25	50	120	115	100	75
not_humanitarian	923	134	261	5	10	25	50	918	913	898	873
other_relevant_information	727	106	205	5	10	25	50	722	717	702	677
requests_or_urgent_needs	55	8	16	5	10	25	50	50	45	30	5
rescue_volunteering_or_donation_effort	991	144	280	5	10	25	50	986	981	966	941
sympathy_and_support	330	48	94	5	10	25	50	325	320	305	280
Total	**5163**	**752**	**1461**	**50**	**100**	**250**	**500**	**5113**	**5063**	**4913**	**4663**

Canada Wildfires 2016

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	74	11	21	5	10	25	50	69	64	49	24
displaced_people_and_evacuations	266	39	75	5	10	25	50	261	256	241	216
infrastructure_and_utility_damage	176	25	50	5	10	25	50	171	166	151	126
not_humanitarian	55	8	16	5	10	25	50	50	45	30	5
other_relevant_information	218	32	61	5	10	25	50	213	208	193	168
requests_or_urgent_needs	14	2	4	5	10	14	14	9	4	0	0
rescue_volunteering_or_donation_effort	653	95	186	5	10	25	50	648	643	628	603
sympathy_and_support	113	16	32	5	10	25	50	108	103	88	63
Total	**1569**	**228**	**445**	**40**	**80**	**189**	**364**	**1529**	**1489**	**1380**	**1205**

Cyclone Idai 2019

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	62	9	18	5	10	25	50	57	52	37	12
displaced_people_and_evacuations	40	6	11	5	10	25	40	35	30	15	0
infrastructure_and_utility_damage	248	36	70	5	10	25	50	243	238	223	198
injured_or_dead_people	303	44	86	5	10	25	50	298	293	278	253
missing_or_found_people	13	2	4	5	10	13	13	8	3	0	0
not_humanitarian	56	8	16	5	10	25	50	51	46	31	6
other_relevant_information	285	41	81	5	10	25	50	280	275	260	235
requests_or_urgent_needs	100	15	28	5	10	25	50	95	90	75	50
rescue_volunteering_or_donation_effort	1308	191	370	5	10	25	50	1303	1298	1283	1258
sympathy_and_support	338	49	95	5	10	25	50	333	328	313	288
Total	**2753**	**401**	**779**	**50**	**100**	**238**	**453**	**2703**	**2653**	**2515**	**2300**

Hurricane Dorian 2019

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	958	140	271	5	10	25	50	953	948	933	908
displaced_people_and_evacuations	561	82	159	5	10	25	50	556	551	536	511
infrastructure_and_utility_damage	571	83	161	5	10	25	50	566	561	546	521
injured_or_dead_people	42	6	12	5	10	25	42	37	32	17	0
not_humanitarian	612	89	173	5	10	25	50	607	602	587	562
other_relevant_information	1011	147	286	5	10	25	50	1006	1001	986	961
requests_or_urgent_needs	125	18	36	5	10	25	50	120	115	100	75
rescue_volunteering_or_donation_effort	691	101	195	5	10	25	50	686	681	666	641
sympathy_and_support	758	110	215	5	10	25	50	753	748	733	708
Total	**5329**	**776**	**1508**	**45**	**90**	**225**	**442**	**5284**	**5239**	**5104**	**4887**

Hurricane Florence 2018

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	917	134	259	5	10	25	50	912	907	892	867
displaced_people_and_evacuations	446	65	126	5	10	25	50	441	436	421	396

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
infrastructure_and_utility_damage	224	33	63	5	10	25	50	219	214	199	174
injured_or_dead_people	208	30	59	5	10	25	50	203	198	183	158
not_humanitarian	742	108	210	5	10	25	50	737	732	717	692
other_relevant_information	445	65	126	5	10	25	50	440	435	420	395
requests_or_urgent_needs	38	5	11	5	10	25	38	33	28	13	0
rescue_volunteering_or_donation_effort	1034	151	293	5	10	25	50	1029	1024	1009	984
sympathy_and_support	330	48	94	5	10	25	50	325	320	305	280
Total	**4384**	**639**	**1241**	**45**	**90**	**225**	**438**	**4339**	**4294**	**4159**	**3946**

Hurricane Harvey 2017

L# = Labeled set, seed 1, budget #   U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	379	55	107	5	10	25	50	374	369	354	329
displaced_people_and_evacuations	482	70	136	5	10	25	50	477	472	457	432
infrastructure_and_utility_damage	852	124	241	5	10	25	50	847	842	827	802
injured_or_dead_people	488	71	139	5	10	25	50	483	478	463	438
not_humanitarian	287	42	81	5	10	25	50	282	277	262	237
other_relevant_information	1237	180	350	5	10	25	50	1232	1227	1212	1187
requests_or_urgent_needs	233	34	66	5	10	25	50	228	223	208	183
rescue_volunteering_or_donation_effort	1976	288	559	5	10	25	50	1971	1966	1951	1926
sympathy_and_support	444	65	126	5	10	25	50	439	434	419	394
Total	**6378**	**929**	**1805**	**45**	**90**	**225**	**450**	**6333**	**6288**	**6153**	**5928**

Hurricane Irma 2017

L# = Labeled set, seed 1, budget #   U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	429	62	122	5	10	25	50	424	419	404	379
displaced_people_and_evacuations	528	77	150	5	10	25	50	523	518	503	478
infrastructure_and_utility_damage	1317	192	372	5	10	25	50	1312	1307	1292	1267
injured_or_dead_people	626	91	177	5	10	25	50	621	616	601	576
not_humanitarian	430	63	122	5	10	25	50	425	420	405	380
other_relevant_information	1651	240	467	5	10	25	50	1646	1641	1626	1601

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
requests_or_urgent_needs	88	13	25	5	10	25	50	83	78	63	38
rescue_volunteering_or_donation_effort	1113	162	315	5	10	25	50	1108	1103	1088	1063
sympathy_and_support	397	58	112	5	10	25	50	392	387	372	347
Total	**6570**	**958**	**1862**	**45**	**90**	**225**	**450**	**6524**	**6480**	**6354**	**6129**

Hurricane Maria 2017

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	154	22	44	5	10	25	50	149	144	129	104
displaced_people_and_evacuations	92	13	26	5	10	25	50	87	82	67	42
infrastructure_and_utility_damage	999	145	283	5	10	25	50	994	989	974	949
injured_or_dead_people	211	31	60	5	10	25	50	206	201	186	161
not_humanitarian	189	28	53	5	10	25	50	184	179	164	139
other_relevant_information	1097	160	311	5	10	25	50	1092	1087	1072	1047
requests_or_urgent_needs	498	72	141	5	10	25	50	493	488	473	448
rescue_volunteering_or_donation_effort	1384	202	391	5	10	25	50	1379	1374	1359	1334
sympathy_and_support	470	69	133	5	10	25	50	465	460	445	420
Total	**5094**	**742**	**1442**	**45**	**90**	**225**	**450**	**5049**	**5004**	**4869**	**4644**

Kaikoura Earthquake 2016

L# = Labeled set, seed 1, budget # U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	345	50	98	5	10	25	50	340	335	320	295
displaced_people_and_evacuations	61	9	17	5	10	25	50	56	51	36	11
infrastructure_and_utility_damage	218	32	62	5	10	25	50	213	208	193	168
injured_or_dead_people	73	11	21	5	10	25	50	68	63	48	23
not_humanitarian	157	23	44	5	10	25	50	152	147	132	107
other_relevant_information	218	32	61	5	10	25	50	213	208	193	168
requests_or_urgent_needs	17	2	5	5	10	17	17	12	7	0	0
rescue_volunteering_or_donation_effort	145	21	41	5	10	25	50	140	135	120	95
sympathy_and_support	302	44	86	5	10	25	50	297	292	277	252
Total	**1536**	**224**	**435**	**45**	**90**	**217**	**417**	**1491**	**1446**	**1319**	**1119**

Kerala Floods 2018

L# = Labeled set, seed 1, budget #   U# = Unlabeled complement, seed 1, budget #

Class Label	Train	Dev	Test	L5	L10	L25	L50	U5	U10	U25	U50
caution_and_advice	97	14	28	5	10	25	50	92	87	72	47
displaced_people_and_evacuations	39	6	11	5	10	25	39	34	29	14	0
infrastructure_and_utility_damage	207	30	59	5	10	25	50	202	197	182	157
injured_or_dead_people	254	37	72	5	10	25	50	249	244	229	204
not_humanitarian	319	47	90	5	10	25	50	314	309	294	269
other_relevant_information	669	97	189	5	10	25	50	664	659	644	619
requests_or_urgent_needs	413	60	117	5	10	25	50	408	403	388	363
rescue_volunteering_or_donation_effort	3005	438	851	5	10	25	50	3000	2995	2980	2955
sympathy_and_support	585	85	165	5	10	25	50	580	575	560	535
Total	**5588**	**814**	**1582**	**45**	**90**	**225**	**439**	**5543**	**5498**	**5363**	**5149**

Interpretation Guide

How to read the tables above and understand their impact on the LG-CoTrain pipeline

How to Read This Table

Column groups

- **Train / Dev / Test** — the full dataset splits for this event (all available samples across all seed sets).
- **L5, L10, L25, L50** — the labeled training subset at the given budget (target: that many samples *per class*), using seed 1 as a representative. This is the set the two BERT models are supervised-trained on in Phase 1 (weight generation) and Phase 3 (fine-tuning).
- **U5, U10, U25, U50** — the unlabeled complement at each budget (seed 1). These are the tweets *excluded* from the labeled set. They are paired with GPT-4o pseudo-labels to form D_{LG}, the sole training set for Phase 2 co-training. A larger labeled budget means a smaller unlabeled complement.

Heat-map colouring — each cell is coloured relative to the largest class count in the Train column for that event:

Blue — high (≥ 60% of max train count)   **Green — medium (30–59%)**   **Yellow — low (10–29%)**   **Red — very low (< 10%)**
Grey — 0 (absent)

Warning Signs to Look For

- **L# < budget for a class** — the class does not have enough samples to fill the requested budget. All available samples are used, but the labeled set becomes imbalanced. *Example: budget = 25 but the class only has 14 training samples → L25 shows 14, not 25.*
- **L# stays the same across multiple budget levels** — the class has hit its natural ceiling; all available samples are already included. Increasing the budget no longer adds real training data for that class. *Example: if both L25 and L50 show 14, the class has exactly 14 samples in the training set.*
- **U# = 0 for a class** — all samples of that class were consumed by the labeled set; none remain for the unlabeled complement D_{LG}. Co-training in Phase 2 receives no genuine examples of this class, only noise from GPT-4o misclassifications.
- **L# = Train count for a class** — the entire training set for that class is labeled. Combined with U# = 0, all available data has been exhausted; the algorithm has no headroom for semi-supervised learning on that class.

Scenario 1 — Unbalanced Labeled Set (some L# < budget)

All three training phases are degraded for the underrepresented class. The core problem is that standard cross-entropy loss treats every sample equally — majority classes dominate the gradient signal because they appear more often per epoch.

- **Phase 1 (Weight Generation) — unreliable lambda weights.** Model 1 and Model 2 are each trained on half the labeled set (D_{l1} and D_{l2}). For a class with only 14 total samples each model sees roughly 7 examples — compared to hundreds for majority classes. With so few examples the model's softmax probability for that class stays low and fluctuates unpredictably across epochs. The *WeightTracker* records these noisy probabilities: *confidence* (mean probability) is low and *variability* (std) is inflated. The resulting lambda weights — which determine how much each D_{LG} sample contributes to Phase 2 training — are either near-zero (the sample is ignored) or erratic (the sample receives inconsistent weight). Neither outcome is useful.
- **Phase 2 (Co-Training) — a self-reinforcing feedback loop.** Lambda weights scale each sample's contribution to the co-training loss. Samples pseudo-labeled as the rare class receive systematically low weights, so they contribute little gradient, so the model does not improve on that class, so the next epoch's weights remain low — a vicious cycle with no internal break. For example, if *rescue_volunteering_or_donation_effort* has 653 training samples and *requests_or_urgent_needs* has only 14, Phase 2 learns an excellent boundary for the former and a weak, uncertain one for the latter, regardless of how many pseudo-labeled examples of the rare class exist in D_{LG}.
- **Phase 3 (Fine-Tuning) — too little data to correct Phase 2 bias.** Fine-tuning revisits only D_{l1} and D_{l2} — the same small labeled set split in half again. Seven genuine examples cannot overcome a poorly calibrated decision boundary built over many co-training epochs. Early stopping compounds the problem: the overall dev macro-F1 may look acceptable because all majority classes improved, causing early stopping to fire before the minority class is properly learned.

> **Example — Canada Wildfires 2016, *requests_or_urgent_needs*:** This class has only 14 samples in Train. At any budget ≥ 25, all 14 are consumed by the labeled set. Each model receives only ~7 examples, compared to ~326 for *rescue_volunteering_or_donation_effort* at the same budget. The macro-F1 contribution from this class is consistently much lower than from majority classes, anchoring the event's overall score below what a balanced dataset would achieve.

Scenario 2 — Unbalanced Unlabeled Set (some U# is very low or zero)

D_{LG} is the *exclusive* training data for Phase 2. Its class composition is therefore critical. Two distinct sub-cases arise.

- **Sub-case A — U# is low but > 0: weak but genuine signal.** Phase 2 still receives real examples of the class with (hopefully) correct pseudo-labels. The lambda weighting partially compensates by amplifying high-confidence samples, but the proportionally small class representation means the model underfits that class relative to majority classes. Performance will be below ideal, but the learning direction is at least correct.
- **Sub-case B — U# = 0: co-training trains on noise, actively corrupting the decision boundary.** When no real samples of class C exist in D_{LG}, the only way class C appears there is through GPT-4o misclassification errors — tweets from other classes that GPT-4o incorrectly tagged as class C. Phase 2 then uses these *false* pseudo-labels as genuine training signal:
  - The cross-entropy loss pushes the model to classify those tweets as class C.
  - Those tweets actually belong to other classes, so the model is learning the wrong feature associations for class C.
  - The decision boundary for class C shifts toward the feature distributions of whatever classes GPT-4o confused with it.

  This is actively harmful — worse than simply ignoring the class. Phase 3 fine-tuning must both relearn the correct boundary *and* fight the corrupted one from Phase 2, armed with only a handful of genuine samples.

> **Example — Canada Wildfires 2016, *requests_or_urgent_needs* at budget 25/50:** U25 = 0 and U50 = 0. No genuine tweets of this class are available for co-training. Any pseudo-labels tagged as *requests_or_urgent_needs* in D_{LG} come from other classes that GPT-4o mislabelled — for instance, a *rescue_volunteering_or_donation_effort* tweet containing "urgent" might be mislabelled. The co-training model then learns to associate "urgent volunteering appeals" with *requests_or_urgent_needs*, corrupting the representation of both classes simultaneously.

Scenario 3 — Both Labeled and Unlabeled Are Unbalanced (the Worst Case)

All three phases reinforce each other's weaknesses. Recovery is impossible for the affected class. This occurs when a class has too few labeled samples (Scenario 1) *and* no real samples in D_{LG} (Scenario 2, sub-case B) simultaneously.

- **Phase 1:** Too few labeled samples → low, noisy probabilities → unreliable lambda weights.
- **Phase 2:** Zero real samples in D_{LG} → trains entirely on GPT-4o misclassification noise → corrupted decision boundary.

- **Phase 3:** Too few labeled samples to correct the corruption accumulated in Phase 2.

The budget paradox — more data can produce lower performance. As the budget grows, the labeled set expands but the unlabeled complement shrinks. For a rare class this creates a non-monotonic performance curve where macro-F1 at budget 25 can be *lower* than at budget 10 for the same event. The table below uses *requests_or_urgent_needs* in Canada Wildfires 2016 (14 total train samples) as a concrete illustration:

Budget	Labeled samples	Real samples in D_{LG}	Co-training signal
L5	5	9	✓ 9 genuine examples available
L10	10	4	~ 4 genuine examples (signal shrinking)
L25	14 (capped)	0	✗ noise only — can be *worse* than L10
L50	14 (capped)	0	✗ noise only — identical situation to L25

More labeled data does not always mean better performance when the unlabeled complement is simultaneously depleted by that increase.

> **Root cause — a violated semi-supervised learning assumption.** LG-CoTrain assumes the unlabeled data distribution reflects the true class distribution. When D_{LG} is constructed by excluding the labeled set and a class is rare enough that the budget ceiling exhausts all of its available samples, this assumption breaks completely for that class. The algorithm cannot distinguish "this class is genuinely rare in the wild" from "this class was artificially removed from the unlabeled pool by the experimental design." The result: the pipeline can actively harm performance on rare classes at higher budgets — a failure mode invisible from the results tables alone, but clearly visible in the data distribution tables above.