In-Domain Encoders Beat Scale: A Stable, Honestly-Calibrated System for Enthymeme Detection and Reconstruction

Muhammad Rafi[1,*], Ajay Kumar[1] and Abdul Rehman Azam[1]

[1]National University of Computer and Emerging Sciences (FAST-NU), Karachi, Pakistan

Abstract

We describe team FAST-NU's system for the MediaEval 2026 task "Missing Pieces and Misinformation," which detects implicit argument components in vaccine and immigration tweets (Task 1: implicit premise, implicit conclusion, or none) and reconstructs the missing proposition as one sentence (Task 2). Our main finding is that small in-domain encoders (BERTweet-COVID) clearly beat the much larger DeBERTa-v3-large, which collapsed to the majority class and, once stabilised, underfit. We submit a soft-vote of three BERTweet-COVID fine-tunes trained with an explicit anti-collapse recipe and an honest, cross-validated decision threshold; a KL model over the five-annotator distribution forms Run 2, and a generate-then-judge pipeline handles Task 2. Our best run reaches 0.678 binary macro-F1 (1A) and 0.483 (1B); disagreement-aware modelling improves macro-F1 and calibration, and Task 2 attains a ROSCOE-SS of 0.902 on matched instances. We also quantify a $\approx$0.71 annotator-disagreement ceiling that no lever we tried could exceed.

1. Introduction

An enthymeme is an argument in which a premise or the conclusion is left unstated, to be supplied by the reader. Such gaps are common in persuasive social media and are where misleading reasoning often hides: the force of a vaccine-hesitant or anti-immigration tweet usually rests on an assumption its author never writes down, and surfacing that assumption is a prerequisite for fact-checking [1]. The MediaEval 2026 task [1] makes this concrete on tweets about COVID-19 vaccination and UK immigration. Task 1 classifies each tweet as *implicit_premise*, *implicit_conclusion*, or *none*, with a binary enthymeme-vs.-none view as the primary 1A metric; Task 2 reconstructs the missing proposition as a single declarative sentence.

Two properties of the data drive every design choice. It is small—1,066 training tweets after the development split, 148 withheld for test—and imbalanced, with about two-thirds *none* and only $\sim$4% *implicit_conclusion*. Both push models toward majority-class collapse, the failure documented by the dataset authors [1]. We initially planned a DeBERTa-v3-large ensemble; it collapsed and destabilised on this data, so we pivoted to a smaller, in-domain, carefully-stabilised system. Our contributions are: (i) the finding that in-domain pretraining beats scale here; (ii) a root-caused negative result on DeBERTa-v3 under gradient checkpointing; (iii) an honest, leakage-free evaluation that exposes a ≈ 0.71 annotator-disagreement ceiling; and (iv) evidence that modelling disagreement improves both macro-F1 and calibration.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
muhammad.rafi@nu.edu.pk (M. Rafi); k230514@nu.edu.pk (A. Kumar); k230061@nu.edu.pk (A. R. Azam)

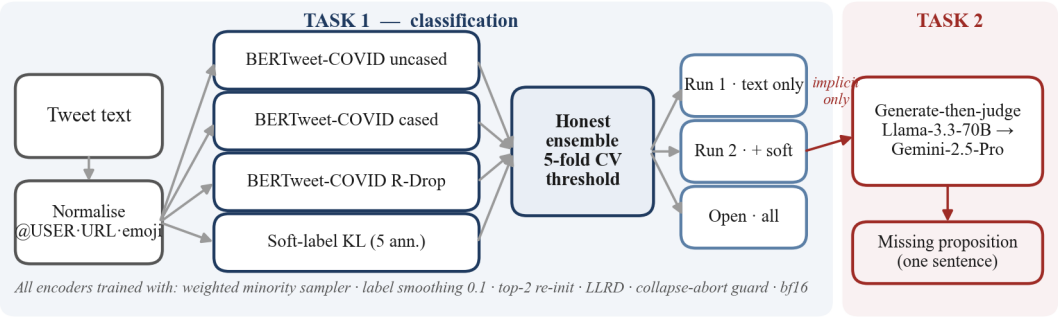

Figure 1: End-to-end FAST-NU system. Four in-domain encoders, trained under a shared anti-collapse recipe, feed an honest cross-validated ensemble; implicit predictions drive the Task 2 pipeline.

2. Related Work

The corpus and task originate with Pastor and Oostdijk [1], who provide the dataset and document the class-collapse failure that motivates our recipe. Our encoders are domain-adaptive: BERTweet [2] is pre-trained on English tweets with the @USER/HTTPURL normalisation we adopt, and COVID-Twitter-BERT [3] specialises it to pandemic discourse; we benchmark these against the larger general encoder DeBERTa-v3 [4]. Bucher and Martini [5] find that fine-tuned compact models still beat zero-shot LLMs on constrained classification, which we confirm. For stable small-data fine-tuning we use top-layer re-initialisation [6], discriminative learning rates [7], and R-Drop [8]. Rather than discard the five annotations per tweet, we treat their distribution as a soft target—an instance of learning from disagreement [9, 10]. For Task 2 we adopt the generate-then-judge paradigm that won CQs-Gen 2025 [11] and evaluate with the task's prescribed ROSCOE [12] and BERTScore [13].

3. Approach

Figure 1 shows the system. Task 1 is an ensemble of in-domain encoders fused by an honest decision rule; Task 2 generates and judges propositions only for tweets predicted to be enthymemes.

3.1. Data and Preprocessing

We use the final release: a stratified 80/20 split at seed 42 gives 1,066 train / 267 dev tweets (test withheld). The majority vote is the hard label; the five-annotator vote vector is the soft target scored by the Task 1B cross-entropy. We normalise each tweet as BERTweet expects (NFKC; URLs → HTTPURL; mentions → @USER; emoji demojised)—our single largest non-architectural lever (+0.05–0.09 binary-F1) [2]. For imbalance we use one lever only: a weighted minority sampler (inverse-frequency exponent 0.7); stacking it with a class-weighted loss destabilised gradients, so we dropped the latter.

3.2. Encoder Roster and Anti-Collapse Recipe

The roster is the base-size BERTweet-COVID family [2, 3]—uncased, cased, and an R-Drop variant [8]—plus a soft-label KL model on the annotator distribution. All share one recipe built against collapse: weighted sampler, label smoothing 0.1, top-2-layer re-initialisation [6], layer-wise learning-rate decay [7], a cosine schedule with warmup, bf16 with a non-finite guard, and a collapse-abort guard that halts any run whose implicit-class F1 stays below 0.05 for three epochs. Each model trains ≤ 10 epochs, early-stopped on dev binary macro-F1 (batch 8, accumulation 4, learning rate $2\mathrm{e}{-5}$).

3.3. Honest Ensemble and Runs

Run 1 uses the three text-only encoders; Run 2 adds the soft-label model so it genuinely uses the annotator labels; the Open Run searches all models. For each run we fuse member probabilities into one $P(\text{enthymeme})$ score and tune the binary threshold *inside* five cross-validation folds, scoring only held-out predictions. This matters: fitting on the same 267 dev rows is ~ 0.04 macro-F1 optimistic, so every dev figure we report is out-of-fold.

3.4. Task 2: Generate-then-Judge

For each enthymeme tweet we generate eight candidates with Llama-3.3-70B—four conditioned on the tweet's Walton scheme at temperature 0.7, four plain at 0.3—and remove duplicates. An independent judge (Gemini-2.5-Pro, with fallbacks) selects the best under four ordered criteria: logical necessity, faithfulness, concreteness, and scheme fit. The generator is constrained to a single declarative sentence (6–35 words) and to the role fixed by the Task 1 label. This separation of generation from selection follows the CQs-Gen 2025 winner [11]; *none* tweets emit an empty proposition.

4. Results and Analysis

4.1. Test Results and the Value of Disagreement

Table 1 gives the official test results and Table 2 the honest dev figures for the shipped runs. The close match between honest dev (0.691) and test (0.672–0.678) confirms the threshold did not overfit. Run 2, which adds the annotator-distribution model, beats text-only Run 1 on every aggregate metric. The hard-label gain is modest (1A $0.6715 \rightarrow 0.6777$; 1B $0.4769 \rightarrow 0.4825$), but the calibration gain is clear (Figure 2a): cross-entropy $0.6834 \rightarrow 0.6576$ (1A) and $0.8795 \rightarrow 0.8477$ (1B), Brier $0.3036 \rightarrow 0.2885$. Since Task 1B is scored partly on the probability vector against soft gold, disagreement-aware modelling is rewarded twice. Figure 2b shows our main weakness: *implicit_conclusion* (six test instances) reaches only 0.14 F1.

4.2. In-Domain Pretraining Beats Scale

We trained roughly twenty encoders on the fixed split; Figure 3 ranks the strongest. The top is entirely Twitter-domain encoders (BERTweet-COVID 0.697, then BERTweet-large, CT-BERT-v2, Twitter-RoBERTa, TwHIN-BERT), while every general-purpose model—and the whole DeBERTa-v3 family—sits at the bottom. On a task this small and domain-specific, in-domain pretraining beats scale, contradicting our initial plan and matching the small-model finding of Bucher and Martini [5].

Table 1
Official test results (Task 1A binary and 1B three-class, Run 1 / Run 2).

Metric	1A R1	1A R2	1B R1	1B R2
macro-F1	0.6715	0.6777	0.4769	0.4825
accuracy	0.6824	0.6892	0.6419	0.6486
cross-entropy	0.6834	0.6576	0.8795	0.8477
Brier	0.3036	0.2885	0.3387	0.3197
F1 none	0.7314	0.7386	0.7314	0.7386
F1 enthymeme	0.6116	0.6167	—	—
F1 premise	—	—	0.5660	0.5660
F1 conclusion	—	—	0.1333	0.1429

Table 2
Honest five-fold CV-threshold development results for the shipped runs.

Run	bin-F1	3cl-F1	soft-CE
Run 1	0.6908	0.4898	0.9585
Run 2	0.6890	0.4905	0.9197
Open	0.6908	0.4898	0.9585

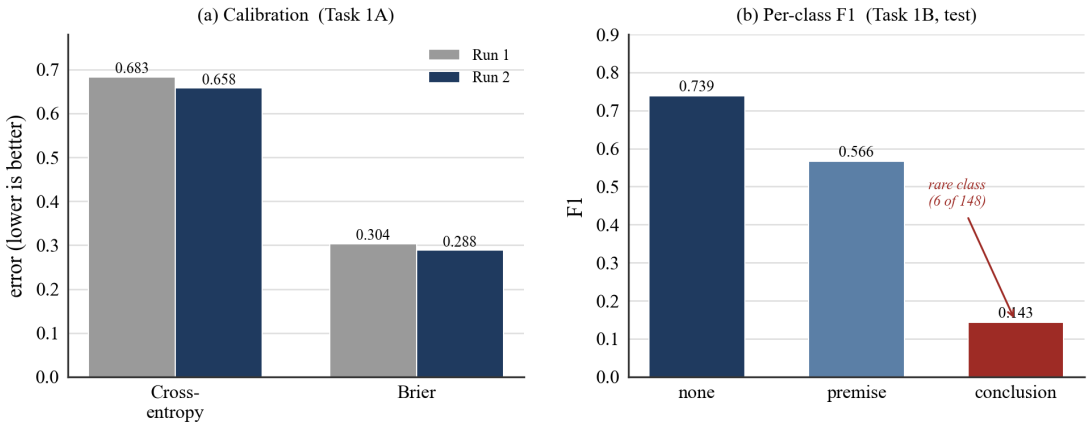

Figure 2: (a) Run 2 (soft labels) improves calibration over Run 1 on Task 1A. (b) Per-class F1 on test (Run 2); the rare *implicit_conclusion* class is the dominant error source.

4.3. A Reproducible DeBERTa-v3 Collapse

In one run, every DeBERTa-v2/v3 model collapsed to ≈0.40 binary-F1 (predicting only *none*) while all non-DeBERTa encoders trained normally. An A/B test traced this to gradient checkpointing on DeBERTa's disentangled attention producing all-NaN logits at the first optimizer step; the trainer's safety guard then skips every batch and the model freezes at initialisation. It is not a mixed-precision artefact (fp32 also produced NaNs). Disabling checkpointing and training in fp32 fixes the NaNs but DeBERTa-v3-large then reaches only 0.557 and underfits, so we dropped it [4, 6].

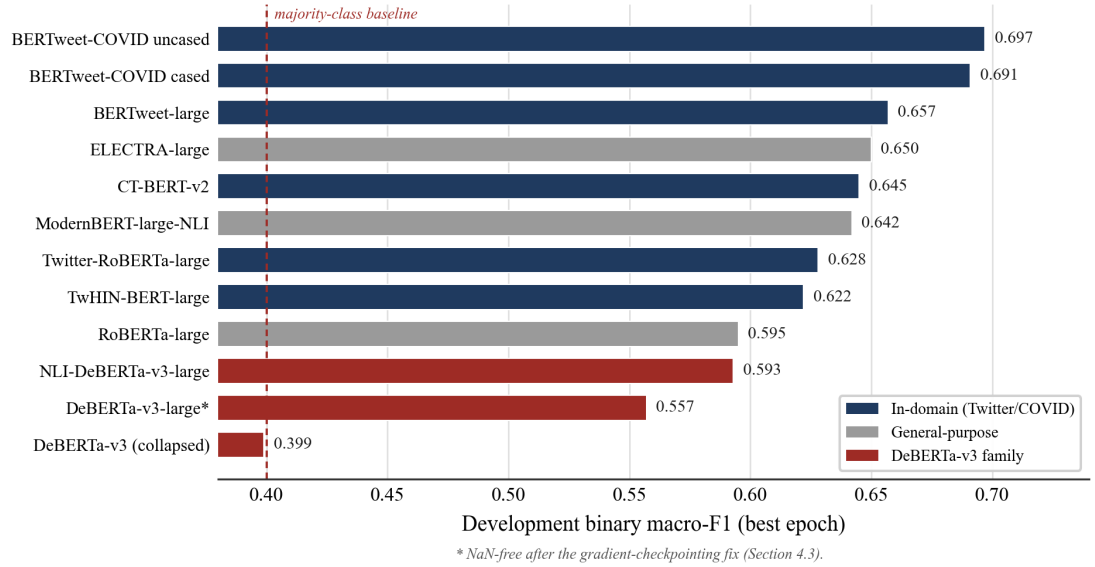

Figure 3: Best dev binary macro-F1 by backbone. In-domain Twitter encoders dominate; the DeBERTa-v3 family collapsed (0.399) or, once stabilised, underfit (0.557).

4.4. An Annotator-Disagreement Ceiling

We could not push binary macro-F1 above 0.73. The honest out-of-fold ceiling is ≈0.706. Larger ensembles, adversarial training, multi-seed averaging, intermediate-task transfer, stacking, and an 11-shot LLM member (0.486) all failed to beat a single in-domain model plus a one-parameter threshold—the members read the same short tweet and are highly correlated, so averaging dilutes rather than decorrelates. The binding constraint is annotator subjectivity (five disagreeing annotators, a 4% conclusion class, 267 dev rows), not capacity; about ten independent methods hit the same wall. Breaking it needs in-domain unlabelled pretraining or more and cleaner labels.

4.5. Task 2

On the subset where both system and annotators produced a reconstruction (n=36), our propositions reach a ROSCOE-SS of 0.9021—high agreement where the pipeline fires. Scored with zeros over the larger set (n=84) it falls to 0.3866; the gap is coverage, not quality, dominated by tweets Task 1 labelled *none* that carried a gold implicit component. Task 2 is therefore gated by Task 1 recall on the implicit classes. For a vaccine-mandate tweet, the system reconstructs the unstated premise as "Forcing people to get vaccinated is a violation of personal freedom and autonomy."

5. Discussion and Outlook

On a small, imbalanced, Twitter-domain task, the highest-leverage moves were domain-matched encoders, an anti-collapse recipe, and honest calibration—not larger models. Our most instructive result is negative: DeBERTa-v3-large was actively harmful here, collapsing then underfitting where small BERTweet-COVID succeeded. Disagreement-aware modelling improved macro-F1 and calibration, and we quantified a ≈ 0.71 ceiling no lever could exceed. The clearest next

steps are in-domain unlabelled pretraining and better recall on the rare *implicit_conclusion* class, which also gates Task 2. We release the system as one reproducible notebook.

Acknowledgments

We thank the organisers, Martial Pastor and Nelleke Oostdijk (Radboud University), for the dataset and evaluation. The corpus contains language hostile toward immigrants and toward vaccination; this is a linguistic-reconstruction task for downstream fact-checking, and the generated propositions do not represent the authors' views.

Declaration on Generative AI

During the preparation of this work, the authors used Anthropic Claude in order to: grammar and spelling checking, paraphrasing and rewording, and improving writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Pastor, N. Oostdijk, A resource for enthymeme detection in controversial political discourse, arXiv preprint arXiv:XXXX.XXXXX (2026).
[2] D. Q. Nguyen, T. Vu, A. T. Nguyen, BERTweet: A pre-trained language model for english tweets, in: Proceedings of EMNLP 2020: System Demonstrations, 2020, pp. 9–14.
[3] M. Müller, M. Salathé, P. E. Kummervold, COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on twitter, arXiv preprint arXiv:2005.07503 (2020).
[4] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, arXiv preprint arXiv:2111.09543 (2021).
[5] M. J. J. Bucher, M. Martini, Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification, arXiv preprint arXiv:2406.08660 (2024).
[6] M. Mosbach, M. Andriushchenko, D. Klakow, On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines, in: International Conference on Learning Representations (ICLR), 2021.
[7] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of ACL 2018, 2018, pp. 328–339.
[8] X. Liang, L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu, R-Drop: Regularized dropout for neural networks, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.
[9] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research (JAIR) 72 (2021) 1385–1470.
[10] A. M. Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics (TACL) 10 (2022) 92–110.
[11] L. Favero, J. A. Pérez-Ortiz, T. Käser, N. Oliver, ELLIS alicante at the CQs-Gen 2025 shared task: Generate-then-select for critical-question generation, in: Proceedings of the 12th Workshop on Argument Mining (ArgMining 2025), 2025.
[12] O. Golovneva, M. Chen, S. Poff, M. Corredor, L. Zettlemoyer, M. Fazel-Zarandi, A. Celikyilmaz, ROSCOE: A suite of metrics for scoring step-by-step reasoning, in: International Conference on Learning Representations (ICLR), 2023.
[13] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: International Conference on Learning Representations (ICLR), 2020.