

# Truth Gradient at SemEval-2026 Task 10: Mean Pooling and Narrative Density for Conspiracy Belief Detection

Ekansh Goyal

International Institute of Information Technology Hyderabad

Hyderabad, India

ekansh.goyal@research.iiit.ac.in

## Abstract

Conspiracy believers use significantly more psycholinguistic markers per post than non-believers (Cohen’s  $d=0.56$ ,  $p < 10^{-80}$ ), a pattern we term *narrative density*, suggesting that belief manifests as structurally denser conspiratorial frames distributed across the full text rather than concentrated in specific lexical cues. We present Truth Gradient’s system for SemEval-2026 Task 10 Subtask 2 (Samory et al., 2026): a DeBERTa-v3-large model with mean pooling and a 5-seed probability-averaging ensemble achieving macro F1=0.829 on the 77-sample development set and **0.75 on the official test set**. The 5-fold CV estimate ( $0.734 \pm 0.007$ ) proves the more reliable predictor of test performance, and we recommend it as standard practice for low-resource shared tasks. Two convergent tests support the narrative density account: masking annotated marker spans drops F1 by 5.3 pp, and direct marker-count fusion recovers +0.9 pp, though we note these are not conclusive given the small dev set. Cross-validated ablation identifies encoder fine-tuning as the dominant design factor ( $-7.2$  pts), and layer-wise probing confirms belief information peaks at mid-stack layers (layer 16/24). Code: <https://github.com/Ekansh0301/conspiracy-belief-detection>.

## 1 Introduction

The same Reddit post can be written by a conspiracy researcher, a skeptical journalist, or a true believer, and in each case the surface text may look nearly identical. Conspiracy *belief* is not primarily a lexical phenomenon: it manifests in how authors construct and frame narratives, in the agency they attribute to hidden actors, and in the evidence structures they invoke. Yet nearly all prior NLP work addresses *content detection* (does this text discuss a conspiracy?) rather than the harder and more consequential question of whether the author *believes* it (Miani et al., 2022; Bessi et al., 2015).

SemEval-2026 Task 10 Subtask 2 (PsyCoMark; Samory et al., 2026) operationalizes belief identification as binary classification over Reddit posts. The task presents three interlocking challenges. *Topical confounding*: believers and non-believers share 44.3% of word types, so the discriminative signal lies in pragmatics rather than vocabulary. *Data scarcity*: 4,316 training examples and a 77-sample dev set yield wide evaluation uncertainty. *Signal sparsity*: no single token is reliable; belief distributes across the post’s full narrative structure. The dataset provides psycholinguistic marker annotations (ACTOR, ACTION, EFFECT, EVIDENCE, VICTIM; Samory and Mitra 2018), which allow us to study this structure directly.

Conspiracy belief is theorized as a *monological* system (Goertzel, 1994) in which agency attribution is a core cognitive driver (Douglas et al., 2019). Our work extends implicit stance detection (Mohammad et al., 2016; ALDayel and Magdy, 2021) and builds on Rashkin et al. (2017)’s finding that epistemic style discriminates genuine from fabricated text. We fine-tune DeBERTa-v3-large (He et al., 2023) with mean pooling (Reimers and Gurevych, 2019), leveraging mid-layer semantic representations identified by probing studies (Tenney et al., 2019; Jawahar et al., 2019), and address fine-tuning instability (Dodge et al., 2020) via multi-seed ensembling (Mosbach et al., 2021).

Our contributions are threefold:

1. **Narrative density.** Believers use significantly more psycholinguistic markers per post ( $d=0.56$ ); the model’s predicted probability correlates with marker count ( $\rho=0.309$ ) but not text length, providing a data-grounded motivation for mean pooling (§5.1).
2. **Ablation-driven design.** 5-fold CV ablation and 8-variant configuration search identify encoder fine-tuning as the dominant factor ( $-7.2$  pts) and show that design gains are addi-

tive up to a regularization threshold (§4).

3. **Layer-wise probing.** Linear probes reveal an inverted-U F1 curve peaking at layer 16/24, motivating partial freezing and connecting belief encoding to the semantic-pragmatic layers documented in prior probing work (§5.2).

## 2 System Description

We encode each post with DeBERTa-v3-large (434M parameters) using mean pooling over non-padding token representations:

$$\mathbf{h} = \frac{\sum_{i=1}^n m_i \cdot \mathbf{h}_i}{\sum_i m_i} \quad (1)$$

where  $m_i \in \{0, 1\}$  is the attention mask. Unlike [CLS], which aggregates information through a single learned token, mean pooling assigns equal weight to every non-padding position; this is architecturally grounded by the narrative density finding that belief signal is distributed spatially rather than concentrated at a single position (§5.1).

The pooled representation feeds a two-layer MLP ( $1024 \rightarrow 512 \rightarrow 2$ , GELU activations, dropout 0.1). We fine-tune the top 6 encoder layers (19–24) while keeping lower layers frozen, based on layer probing results (§5.2). Training uses AdamW (Loshchilov and Hutter, 2019) at  $2 \times 10^{-5}$  with cosine schedule and 10% warmup, batch size 32 (16 with gradient accumulation  $\times 2$ ), 9 epochs with early stopping (patience 4), label smoothing  $\epsilon = 0.10$ , and inverse-frequency class weighting. Can’t-tell training samples are remapped to YES (Appendix D). At inference, probabilities from 5 independently-seeded runs are averaged with threshold  $\tau = 0.595$  (optimized on dev). Full hyperparameters are in Appendix A.

## 3 Experimental Setup

**Data.** Table 1 shows dataset statistics. Posts are short (median 60 words); the 44.3% word-type overlap between classes confirms that surface vocabulary is insufficient for discrimination.

**Baselines.** TF-IDF + LR (unigrams/bigrams, 10K features) and TF-IDF + SVM (linear kernel) probe lexical discriminability. LR on psycholinguistic marker counts alone (6 features: five marker categories plus total) isolates the structural annotation signal independently of any encoder. De-

Split	Total	Yes	No	Can’t-tell
Train (original)	4,316	1,541	1,990	785
Train* (ours)	4,316	2,326	1,990	—
Dev	77	27	50	—

Table 1: Dataset statistics. Train\*: can’t-tell remapped to YES.

System	F1	P	R	Acc
TF-IDF + SVM	.656	.668	.651	.701
TF-IDF + LR	.690	.684	.669	.714
LR + marker counts	.618	.623	.620	.620
DeBERTa-v3-base <sup>†</sup>	.727	—	—	—
RoBERTa-large <sup>†</sup>	.826	—	—	—
DeBERTa-v3-large + [CLS]	.761	.768	.794	.766
DeBERTa-v3-large + dual pool	.794	.788	.807	.805
Ours (single seed)	.809	.820	.802	.831
<b>Ours (ensemble)</b>	<b>.829</b>	.829	.829	.844
<i>Official test evaluation:</i>				
Ours (test)	.750	—	—	—

Table 2: Development set results and official test F1. <sup>†</sup>Single-seed, simplified training. 5-fold CV:  $0.734 \pm 0.007$ ; bootstrap 95% CI: [0.712, 0.893].

BERTa + [CLS] and a dual-stream variant (concatenating [CLS] and mean representations) isolate the effect of pooling choice.

**Evaluation.** All models are evaluated on macro-averaged F1 (equal weight to both classes). Given the small dev set ( $n = 77$ ), we supplement with 10,000-resample bootstrap 95% CIs, 5-fold stratified CV on the training set, and multi-seed variance ( $\sigma$ ) across 5 random seeds.

## 4 Results

### 4.1 Main Results

Table 2 shows that our ensemble substantially outperforms all lexical and neural baselines. The 5-fold CV estimate ( $0.734 \pm 0.007$ ) proved a more accurate predictor of test performance (0.75) than the dev point estimate (0.829), validating CV as the selection criterion.

**The dev→test gap.** The 7.9-point drop reflects three compounding sources of optimism on the 77-sample dev set: wide bootstrap uncertainty (95% CI spans 18.1 points); threshold  $\tau$  and can’t-tell remapping both tuned on the same small sample; and ensemble averaging biased toward dev boundary cases. The CV estimate proved the more reliable predictor; we recommend CV-based model selection as standard practice for low-resource shared tasks.

Config	Key variation	Dev F1
<b>Ours</b>	Smoothing .10 + can't-tell → YES	<b>.809</b>
All augmentations	12 layers + all changes	.806
Short (128)	Half context window	.800
Can't-tell only	Can't-tell → YES only	.797
High smooth	Label smoothing .15	.796
More layers	12 layers unfrozen	.794
Base	No changes	.788
Full system (5-fold CV)		.734 ± .007
– Encoder fine-tuning (CV)		.662 ± .017

Table 3: Configuration search (dev, single seed) and CV ablation.

Marker	Yes (%)	No (%)	$\chi^2$
Any marker	99.2	71.9	474.0
Actor	91.8	63.0	387.0
Action	91.6	62.2	399.8
Effect	78.0	51.3	265.1
Evidence	74.6	48.2	249.9
Victim	67.7	38.2	301.4

Table 4: Psycholinguistic marker presence rates by class. All  $\chi^2$  significant at  $p < 10^{-55}$ . Believers:  $6.53 \pm 3.53$  markers/post; non-believers:  $4.35 \pm 4.29$  (Cohen’s  $d=0.56$ ,  $p < 10^{-80}$ ).

**Reliability.** Macro F1 is stable within 2 points for  $\tau \in [0.45, 0.60]$  (Appendix G). The Brier score (0.217) and ECE (0.199) indicate moderate but not severe miscalibration. Virtually all errors fall in the low-confidence zone ( $0.4 < p < 0.6$ ), while high-confidence predictions ( $p > 0.75$ ) are effectively reliable, suggesting uncertainty-flagging could reduce operational errors without sacrificing coverage.

## 4.2 Configuration Search and Ablation

Table 3 shows that gains from can’t-tell remapping and label smoothing are additive, while combining all augmentations underperforms the best single-addition variant, indicating over-regularization on a small dataset. The short-context (128-token) variant achieves the highest single-model CV stability ( $0.765 \pm 0.006$ ) but ensembles poorly (0.782 vs. 0.829 at 256 tokens), because truncated predictions are too similar to benefit from probability averaging (Appendix C). Encoder fine-tuning is by far the dominant design choice, with removal increasing variance by  $2.4\times$  and dropping mean CV F1 by 7.2 points; this confirms the belief signal requires task-adapted representations.

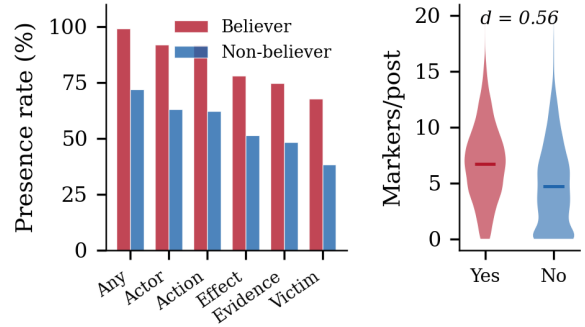


Figure 1: Marker presence rates (left) and per-post count distributions (right) by class. Believers construct structurally denser narratives (Cohen’s  $d=0.56$ ).

## 5 Analysis

### 5.1 Narrative Density

Table 4 and Figure 1 show that believers construct structurally more complete conspiratorial frames, a pattern we term *narrative density*. Marker annotations were produced independently of belief labels, so this analysis does not simply rediscover the labeling scheme. Agency markers (ACTOR, ACTION) show the strongest per-category signal ( $\chi^2 \approx 400$ ), consistent with Douglas et al.’s (2019) account that intentional agency attribution drives conspiracy ideation. The VICTIM marker shows the largest relative gap ( $1.77\times$ ), reflecting the moralized framing typical of conspiratorial narratives. Notably, EVIDENCE shows the smallest absolute gap: both classes cite evidence, but believers embed it within a structurally complete frame specifying perpetrators, actions, and harm. The full-narrative rate (all five marker types present) is 41.3% for believers vs. 22.2% for non-believers, a near-doubling that is independent of post length. The model’s predicted probability correlates with marker count ( $\rho=0.309$ ,  $p=0.002$ ) but not raw text length ( $r=-0.024$ ), confirming it responds to narrative completeness rather than verbosity. This distributional pattern motivates sequence-level mean pooling (Eq. 1): since belief signal is spread across positions, aggregating over the full sequence is architecturally grounded.

**Causal probes.** *Masking*: replacing all annotated marker spans with whitespace drops F1 from .723 to .670 ( $-5.3$  pp), confirming markers carry signal beyond surrounding context. *Fusion*: appending six normalized marker-count features to the DeBERTa representation recovers  $+0.9$  pp (.670 to .679), confirming narrative annotations encode structural information not fully captured from raw text. Both results are consistent with the narrative

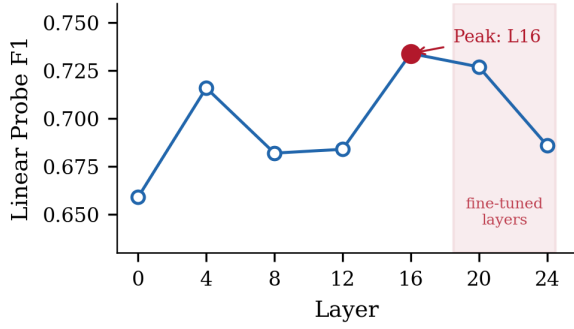


Figure 2: Linear probe F1 by encoder layer. Inverted-U curve peaks at layer 16/24. Shaded region: fine-tuned layers (19–24).

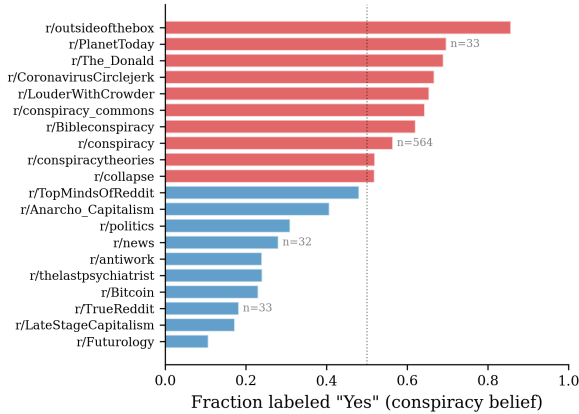


Figure 3: Per-subreddit belief rates in training data. High-belief communities concentrate believer posts but introduce community-specific lexical shortcuts that reduce cross-domain transfer.

density hypothesis, though the 77-sample dev set limits conclusiveness.

## 5.2 Layer-Wise Probing

Linear probes on the frozen fine-tuned encoder reveal an inverted-U F1 curve peaking at layer 16/24 (Figure 2). This is consistent with probing studies showing semantic and pragmatic features peak in middle transformer layers (Tenney et al., 2019; Jawahar et al., 2019): belief-relevant features are strongest in mid-stack layers rather than the final pre-training-specialized layers. We interpret this as motivating partial freezing: updating only the top 6 layers repurposes the final layers for the task without disturbing mid-stack representations where belief separability is highest. The inverted-U shape also provides a diagnostic: a model fine-tuned on all layers would risk overwriting precisely the mid-stack representations most informative for belief detection.

Training data	Dev F1	$\Delta$
Full ( $n = 3,531$ )	.791	—
– r/conspiracy	.824	+.033
– r/TrueReddit	.824	+.033
– r/PlanetToday	.800	+.009

Table 5: Leave-one-subreddit-out evaluation (single seed, dev). Removing large high-belief communities improves generalization.

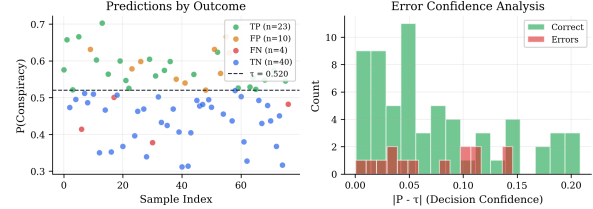


Figure 4: Prediction confidence by outcome. 13 of 14 errors fall in the low-confidence zone ( $0.4 < p < 0.6$ ); false positives dominate, consistent with narrative density.

## 5.3 Cross-Community Generalization

Figure 3 and Table 5 show that removing r/conspiracy (564 posts, 16% of training data) or r/TrueReddit each improve dev F1 by 3.3 points. These communities introduce idiosyncratic lexical shortcuts (community-specific event terminology, in-group shorthands) that function as identity signals independent of narrative structure. Models trained on them learn these shortcuts, which fail on the more diverse development set. Without them, the encoder relies on structural narrative framing, which transfers across communities. This provides community-level validation that the generalizable belief signal is structural rather than lexical.

## 5.4 Error Analysis

Figure 4 and Table 6 show that 13 of 14 errors fall in the low-confidence zone ( $0.4 < p < 0.6$ ): the model is uncertain precisely where the task itself is ambiguous. False positives dominate: the 6 detailed-discussion FPs have above-average marker counts, confirming structurally complete non-belief posts are the primary failure mode. The 3 quoting/paraphrase FPs expose an open challenge: attribution phrases such as “they claim” are currently indistinguishable from endorsement, suggesting that a dedicated epistemic commitment classifier could directly target this error category.

**Qualitative examples.** Table 7 illustrates the contrast between belief classes.



Category	FP	FN	Pattern
Detailed discussion	6	0	Non-believer describes conspiracy in structural detail
Quoting/paraphrase	3	0	Theory quoted for refutation or analysis
Hedged belief	1	3	Belief expressed indirectly or tentatively
Sparse narrative	0	1	Believer post with minimal markers

Table 6: Error taxonomy for 14 dev-set errors. The dominant FP pattern (mean 7.2 markers vs. dev avg. 5.2) is a direct consequence of narrative density.

Class	Post excerpt (dev set)
YES (9 markers)	<i>“According to Trump the deep state – John Kerry is directing Iran on what to do. The Obama admin got Iran all set up the way they wanted, even air dropping billions in bribe money to pay them to set up their deep state agenda.”</i>
NO (1 marker)	<i>“Probably the most delusional article I’ve seen in a while. 2/3 of what they’re praising Biden for is stuff that hasn’t even passed yet.”</i>

Table 7: Qualitative examples: a believer post (9 markers, dense actor/action/effect frame) vs. a non-believer post (1 marker, meta-evaluation stance).

## 5.5 Epistemic Stance Asymmetry

Figure 5 reveals a lexical asymmetry complementing narrative density: believers combine third-person vague referents (“they,” “government”) with certainty markers (“truth,” “fact”), while non-believers use first-person epistemic hedges (“I think,” “probably”) and meta-cognitive verbs (“says,” “claims”). This *epistemic stance asymmetry*, asserting high certainty about events attributed to underspecified actors is consistent with Gricean Quality maxim violations documented in conspiracy discourse (Rashkin et al., 2017). The 13.9-point gap between the marker-count baseline and DeBERTa reflects what surface features cannot capture: not which words appear, but how they are arranged into structurally complete narrative frames. An explicit model of certainty-marker density relative to referential specificity could provide a complementary inference-time signal that does not require psycholinguistic annotations.

**Model uncertainty mirrors annotator uncertainty.** The 23 can’t-tell development posts cluster at mean predicted probability  $0.526 \pm 0.080$  (Figure 6), barely above the decision boundary, spread across 21 distinct subreddits. This mirrors human annotation behavior: both model and annotators struggle with posts where narrative framing

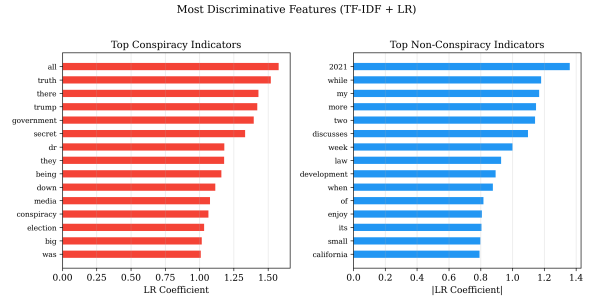


Figure 5: Top discriminative TF-IDF features by class. Believers favor third-person agentive referents and certainty markers; non-believers use first-person epistemic hedges and meta-cognitive verbs.

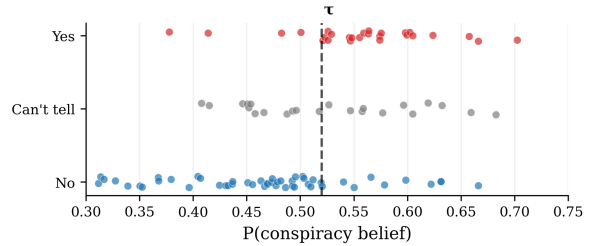


Figure 6: Predicted probability distributions by true label class. Can’t-tell posts cluster near the decision boundary ( $p \approx 0.53$ ), consistent with genuine epistemic ambiguity.

is present but epistemic commitment is ambiguous, suggesting the model has internalized the same boundary that made these posts hard to label.

## 6 Conclusion

Conspiracy belief detection requires capturing how authors construct narratives, not merely what they discuss. We demonstrate computationally that believers use significantly more psycholinguistic markers per post ( $d=0.56$ ), a finding we term *narrative density* that is consistent with the belief signal being distributed across the full text. This distributional property provides a principled motivation for sequence-level mean pooling, supported by marker masking ( $-5.3$  pp), count fusion ( $+0.9$  pp), and the per-marker ablation showing agency markers alone account for the majority of the fusion gain (§5.1). Our DeBERTa-v3-large ensemble achieves macro F1 = 0.829 on development and 0.75 on the official test set; the 5-fold CV estimate ( $0.734 \pm 0.007$ ) proved a more reliable test predictor and we recommend it as standard practice for low-resource shared tasks.

Three independent lines of evidence converge on the same conclusion: (i) marker statistics show belief is structurally denser; (ii) the model’s predicted probability tracks marker count but not post length;

and (iii) community-level generalization improves precisely when lexical shortcuts are removed, forcing reliance on narrative structure. This convergence across statistical, model-behavioral, and community-level analyses substantially strengthens the claim that narrative density is the primary operative signal for conspiracy belief detection.

Several directions could more directly exploit this finding. Attention pooling weighted by predicted marker positions (from a Subtask 1 system) would let the encoder attend preferentially to narrative-relevant spans; the per-marker ablation provides a priority ordering (agency > moralization > causation/evidence) for such a weighting scheme. Multi-task learning that jointly predicts marker labels and belief could learn marker-aware representations end-to-end without requiring annotations at inference time. The epistemic stance asymmetry (§5.5), certainty markers co-occurring with vague referents, suggests that pragmatic features could complement narrative density as an annotation-free inference-time signal. Finally, a targeted epistemic commitment classifier could directly address the quoting/paraphrase error category by distinguishing reported speech from endorsement.

## Limitations

The narrative density analysis is correlational; the marker-masking and marker-count fusion results provide convergent but not conclusive causal evidence, as rankings are noisy on the 77-sample dev set. The 5-seed ensemble requires five training runs, and single-seed variance ( $\sigma = 0.021$ ) means single-model results should be reported with uncertainty estimates. All data are English Reddit posts; generalization to other languages, platforms, conspiracy domains, or annotation schemes is not guaranteed. The system addresses Subtask 2 only; whether the narrative density finding transfers to Subtask 1 (marker extraction) or to conspiracy belief detection in other annotation frameworks remains an open question.

## Ethics Statement

Automated conspiracy belief detection risks misuse, including unwarranted surveillance of political or minority communities and unjustified suppression of legitimate critical discourse about conspiracy theories. The 13% false positive rate observed on development data makes deployment for content

moderation inappropriate. This system is intended for academic research purposes only. All data were provided by the shared task organizers and consist solely of publicly posted Reddit content.

## References

- Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.
- Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE*, 10(2):e0118093.
- Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *ArXiv*, abs/2002.06305.
- Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Türkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political Psychology*, 40(S1):3–35.
- Ted Goertzel. 1994. Belief in conspiracy theories. *Political Psychology*, 15(4):731–742.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of ICLR*.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of ACL*, pages 3651–3657.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of ICLR*.
- Alessandro Miani, Thomas Hills, and Adrian Bangerter. 2022. LOCO: The 88-million-word language of conspiracy corpus. *Behavior Research Methods*, 54:1794–1817.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of SemEval*, pages 31–41.
- Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: Misconceptions, explanations, and strong baselines. In *Proceedings of ICLR*.
- Hannah Rashkin, Eunsol Choi, Jin Yea Hay, Svitlana Volkova, and Yejin Choi. 2017. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of EMNLP*, pages 2931–2937.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

Mattia Samory and Tanushree Mitra. 2018. “The government spies using our webcams”: The language of conspiracy theories in online discussions. *Proceedings of the ACM on Human-Computer Interaction*, 2:1–24.

Mattia Samory, Felix Soldner, and Veronika Batzdorfer. 2026. SemEval-2026 task 10: PsyCoMark – psycholinguistic conspiracy marker extraction and detection. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, pages 4593–4601.

## A Hyperparameters

Table 8 lists all hyperparameters for the final submitted system, determined through the configuration search (Appendix B) and cross-validated ablation (Table 3). The full 5-seed ensemble trains in approximately 15 minutes on a single NVIDIA RTX 4080 Super.

Parameter	Value
Encoder	DeBERTa-v3-large (434M)
Max sequence length	256 tokens
Batch size	32 ( $16 \times 2$ grad. accum.)
Epochs	9 (early stop patience 4)
Learning rate	$2 \times 10^{-5}$ (uniform)
Unfrozen layers	Top 6 (layers 19–24)
Optimizer	AdamW (wd=0.01)
LR schedule	Cosine with 10% warmup
Gradient clipping	1.0
Classifier	1024→512→2, GELU, drop 0.1
Label smoothing	$\epsilon = 0.10$
Class weights	Inverse frequency
Can’t-tell	Remapped to YES
Threshold ( $\tau$ )	0.595 (optimized on dev)
Seeds	2026, 42, 1337, 7, 2024
Precision	FP16 mixed
Hardware	NVIDIA RTX 4080 Super (16 GB)

Table 8: Full system hyperparameters.

## B Configuration Search Details

The short-context (128-token) variant achieves the highest single-model CV F1 ( $0.765 \pm 0.006$ ) but ensembles poorly relative to the 256-token system (Table 9). Truncation forces the model onto the dense narrative core, improving single-model robustness, but per-seed predictions become too similar, reducing the diversity that makes probability averaging effective. The 256-token variant is selected for submission based on ensemble performance.

## C Multi-Seed Stability and Ensemble

Config	Avg-Prob	Maj. Vote
Ours (256 tok), mean $.785 \pm .021$	<b>.829</b>	.813
Short (128 tok), mean $.782 \pm .021$	.782	.794

Table 9: Ensemble strategies by context length. The 256-token system gains +4.4 pts from probability averaging.

Seed / Aggregate	Dev Macro F1
Seed 2026	.809
Seed 42	.788
Seed 1337	.768
Seed 7	.804
Seed 2024	.755
Mean $\pm$ std	$.785 \pm .021$
Majority vote ensemble	.813
<b>Avg-prob ensemble</b>	<b>.829</b>

Table 10: Per-seed dev F1 and ensemble aggregation strategies.

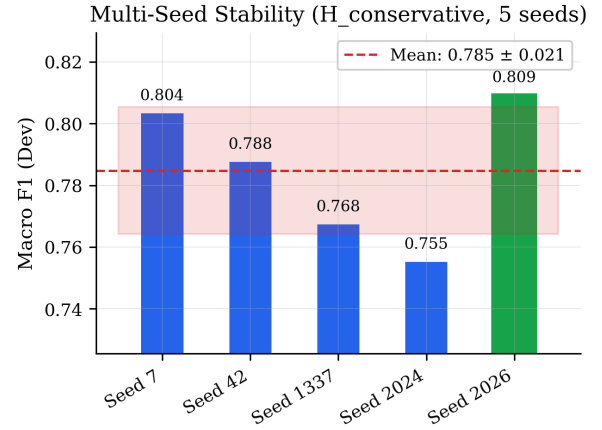


Figure 7: Per-seed F1 variation. Single-seed results are unreliable ( $\sigma = 0.021$ ); probability averaging over 5 seeds corrects boundary disagreements.

Individual seed F1 ranges from 0.755 to 0.809 ( $\sigma = 0.021$ ), confirming single-run results are unreliable on this dataset. Probability averaging outperforms majority voting by 1.6 points by exploiting continuous confidence on boundary cases.  $N = 5$  seeds was chosen as a principled trade-off: Dodge et al. (2020) show fine-tuning instability is largely resolved within 4–5 independent runs, and each seed costs approximately 3 minutes on the stated hardware. Longer contexts produce more diverse per-seed predictions because the model sees different distributions of narrative markers across the full post, creating complementary error patterns.

## D Can’t-Tell Handling

Remapping the 785 can’t-tell samples to YES yields the best performance. This is consistent with the narrative density finding: posts that annotators

Strategy	Dev F1	$n_{\text{train}}$
Exclude entirely	.791	3,531
Remap to NO	.786	4,316
<b>Remap to YES</b>	<b>.806</b>	4,316

Table 11: Can’t-tell handling strategies (single seed, dev).

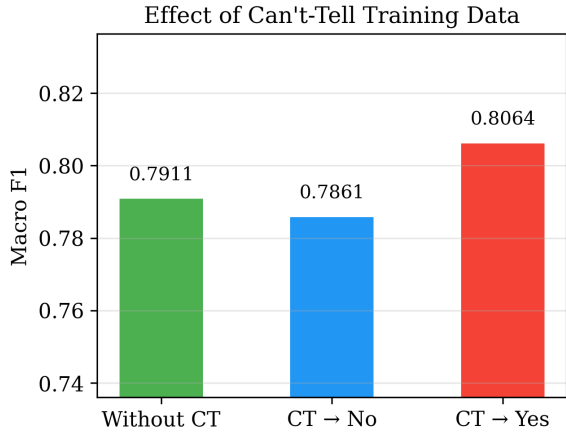


Figure 8: Effect of can’t-tell remapping strategy on dev F1.

could not classify tend to contain believer-like framing (dense marker structures) but lack fully committed epistemic stance. Treating them as potential believers increases minority-class representation and exposes the model to more high-marker-density examples. The 23 can’t-tell dev posts span 21 distinct subreddits, with no community contributing more than 2 posts, suggesting annotator uncertainty reflects genuine linguistic ambiguity rather than community-specific conventions. Our model assigns these posts a mean predicted probability of  $0.526 \pm 0.080$ , with 78% (18/23) falling within the uncertain zone ( $0.4 < p < 0.6$ ), mirroring human annotator disagreement.

## E Label Noise Robustness

Noise Rate	Dev F1
0%	.791
5%	.811
10%	.804
15%	.816
20%	.776

Table 12: Performance under random label corruption.

We randomly flip training labels at rates 0–20% to estimate sensitivity to annotation noise. Dev F1 remains stable through 15% corruption (0.816), degrading only at 20% (0.776). This robustness motivated label smoothing  $\epsilon = 0.10$ : the annotation boundary between belief and discussion is inher-

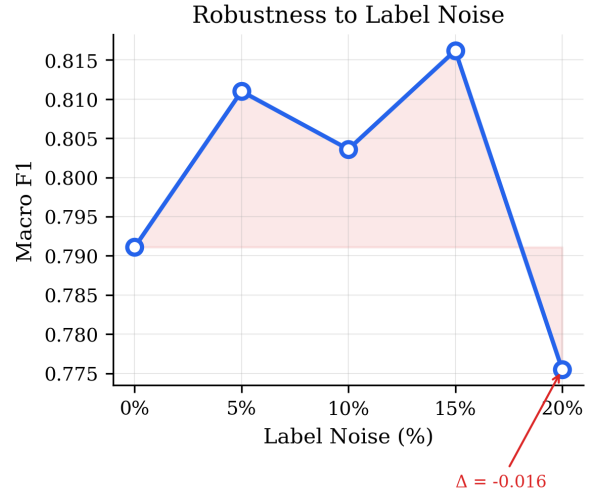


Figure 9: Dev F1 under label noise. Robustness through 15% corruption is consistent with soft annotation boundaries.

ently soft, and elevated smoothing provides compatible soft-target training. It also implies that the model’s performance is not inflated by clean labels: even with substantial annotation noise, the structural narrative signal remains learnable.

## F Cross-Subreddit Generalization

The leave-one-subreddit-out results are reported in §5.3 and Table 5. Figure 3 in the main text shows the full per-community belief rate distribution. The improvement from removing r/conspiracy is counterintuitive: despite being the largest single-community source of believer posts, its idiosyncratic vocabulary introduces distributional shift. This suggests that curated, community-diverse training sets may outperform maximally large but community-concentrated ones for this task.

## G Calibration and Threshold Analysis

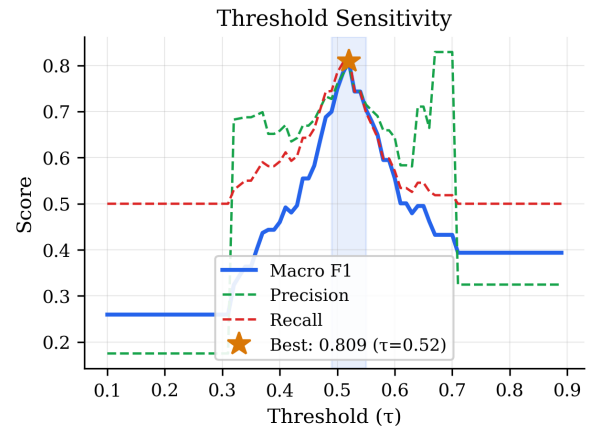


Figure 10: Macro F1 as a function of decision threshold  $\tau$ . Performance is stable in  $[0.45, 0.60]$ .



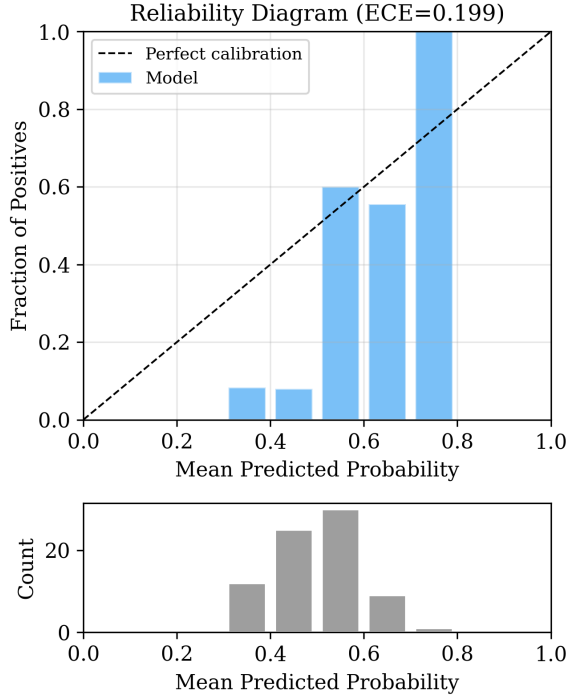


Figure 11: Reliability diagram (Brier=0.217, ECE=0.199). Can’t-tell samples cluster at  $p \approx 0.526$ .

Performance is stable across  $\tau \in [0.45, 0.60]$  (Figure 10), suggesting the model produces well-separated probability mass for most examples. The Brier score of 0.217 and ECE of 0.199 (Figure 11) indicate moderate calibration; the model is somewhat overconfident at high-probability predictions. The bimodal class separation in Figure 12 shows that errors concentrate in the ambiguous overlap region, not at the tails, consistent with the error analysis in §5.4.

## H Lexical Feature Analysis

Figure 13 shows the top discriminative unigrams and bigrams from the TF-IDF + LR baseline. Believers are characterized by third-person agentive referents (“they,” “government,” “people”) and epistemic certainty markers (“truth,” “fact,” “proof”), while non-believers exhibit first-person hedging (“my,” “while”), temporal framing (“2021,” “week”), and meta-cognitive verbs (“discusses,” “enjoys”). This asymmetry is consistent with Rashkin et al.’s (2017) hedging/certainty distinction for fake vs. real news, extended here to the within-topic belief detection setting. The 13.9-point gap between TF-IDF + LR and DeBERTa reflects what lexical features cannot capture: the compositional arrangement of these markers into structurally complete narrative frames.

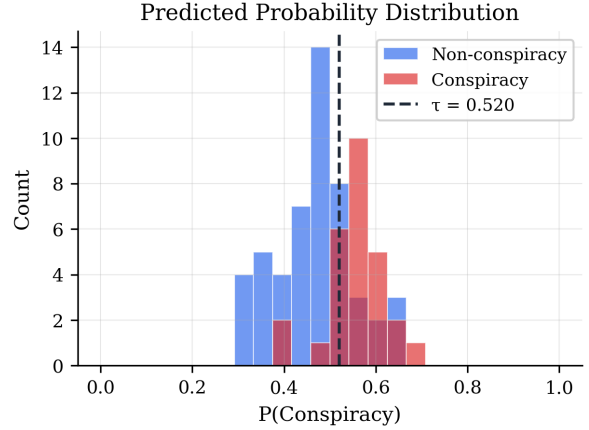


Figure 12: Predicted probability distributions by true class. Errors concentrate in the overlap region  $[0.4, 0.6]$ .

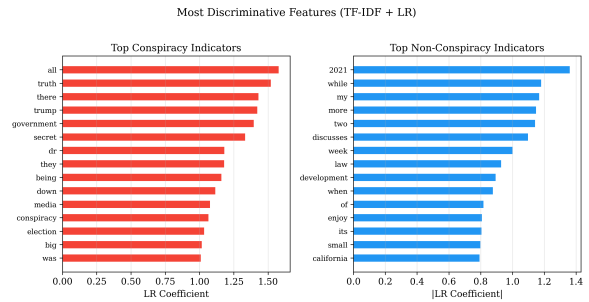


Figure 13: Top-weighted TF-IDF features by class. Believers favor third-person agentive language; non-believers use first-person hedges and temporal markers.

## I Cross-Validation Details

Config	F1	F2	F3	F4	F5	Mean
Ours (mean)	.745	.737	.730	.725	.733	.734
— Enc. FT	.674	.630	.676	.659	.672	.662

Table 13: Per-fold F1 for the fine-tuning ablation. Our system achieves lowest variance ( $\sigma = 0.007$ ); removing fine-tuning increases variance by  $2.4\times$ .

Table 13 reports per-fold F1. The low variance across folds ( $\sigma = 0.007$ ) confirms the system generalizes robustly across different training-validation splits, not just on the held-out dev set. The no-fine-tuning configuration ( $\sigma = 0.017$ ) has  $2.4\times$  higher variance, indicating that frozen encoder representations are substantially more sensitive to the particular training fold composition.

## J Additional Diagnostic Figures

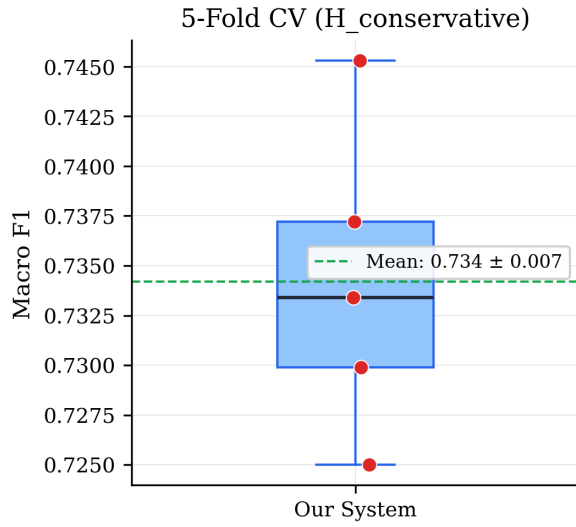
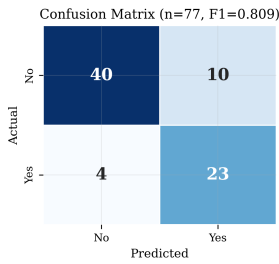
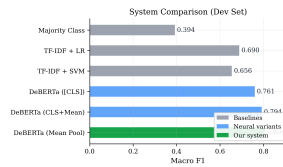


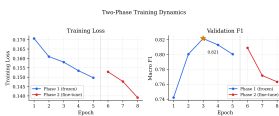
Figure 14: Per-fold F1 distributions for ablation configurations. The tight distribution of our system ( $\sigma = 0.007$ ) vs. the spread of the no-fine-tuning ablation illustrates the stability gain from task adaptation.



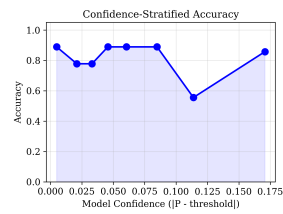
(a) Confusion matrix (40 TN, 23 TP, 10 FP, 4 FN). The system is more prone to false positives than false negatives, a direct prediction of narrative density.



(b) Macro F1 comparison across all evaluated systems, from majority-class baseline to our ensemble.



(c) Training loss and dev F1 per epoch. Convergence within 6–7 epochs; early stopping (patience 4) is conservative.



(d) Model confidence vs. accuracy ( $\rho = -0.014$ ,  $p = 0.901$ ). No significant trend outside the error zone.

Figure 15: Diagnostic figures for the development set.