

Confidence Calibration and Semantic Error Analysis for Ambiguous Coreference Resolution in User-Generated Social Media Text

Yuewei Yuan

CailynLamastus01@outlook.com

University of Pennsylvania Philadelphia, United States 19104

Jialei Huang

Columbia University New York, United States 10027

Jiayang Yin

Johns Hopkins University Baltimore, United States 21218

Tianyi Xu

Georgetown University Washington, D.C., United States 20037

Research Article

Keywords: ambiguous coreference, coreference resolution, heterogeneous relational graph, confidence calibration, semantic error

Posted Date: September 4th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10900526/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Confidence Calibration and Semantic Error Analysis for Ambiguous Coreference Resolution in User-Generated Social Media Text

Yuewei Yuan*

University of Pennsylvania
Philadelphia, United States 19104
Corresponding author: yuewei@alumni.upenn.edu

Jiayang Yin

Johns Hopkins University
Baltimore, United States 21218
yyin33@jh.edu

Jialei Huang

Columbia University
New York, United States 10027
jh3882@columbia.edu

Tianyi Xu

Georgetown University
Washington, D.C., United States 20037
tx52@georgetown.edu

Abstract: *User-generated social media text is brief, elliptical, and structurally irregular, making ambiguous coreference difficult to resolve reliably. We construct a primarily synthetic dataset of 86,400 text fragments, 214,700 entity mentions, and 67,300 coreferential expressions using controlled templates and manual validation. AmbiGraph-Coref represents referential expressions, candidate entities, comment chains, and contextual fragments in a heterogeneous relational graph, and combines pretrained language representations, graph propagation, candidate ranking, multitask learning, confidence calibration, and an explicit UNRESOLVABLE mechanism. On the test set, the model achieves an F1 score of 0.856, accuracy of 0.842, and MRR of 0.887; joint calibration reduces ECE from 0.184 to 0.057 and lowers high-confidence errors. Downstream evaluation further indicates that calibrated reference uncertainty reduces incorrect semantic binding under the constructed setting. The results support reliability-aware coreference resolution for ambiguous social-media text while highlighting the need for validation on authentic real-world social-media platform data across broader interaction conditions and evidence levels.*

Keywords: *ambiguous coreference; coreference resolution; heterogeneous relational graph; confidence calibration; semantic error*

1 Introduction

Coreference resolution has progressed from rule-based matching to neural and generative modeling. Bohnet et al. [1] proposed a seq2seq transition-based system; later studies improved character-level consistency [2], multilingual evaluation and transfer [3]–[6], mention representation [7], and text-generation formulations [8], [9]. These advances strengthen antecedent prediction and representation learning, but generally emphasize whether a selected antecedent is correct rather than whether its confidence is reliable.

Confidence calibration and selective prediction are important for ambiguous social-media text because predicted probabilities may remain poorly aligned with empirical correctness when inputs are incomplete or corrupted [10]. In short replies, ellipsis, cross-level shifts, omitted subjects, and competing entities can leave several antecedents plausible or leave no supported antecedent, yet conventional softmax systems may still return a high-confidence Top-1 choice.

Existing systems therefore insufficiently distinguish competition among plausible candidates from the absence of adequate evidence. A forced entity binding can propagate semantic errors to sentiment, risk identification, and other downstream tasks, so candidate-level confidence and abstention require explicit modeling.

AmbiGraph-Coref addresses this gap with heterogeneous relational graphs, candidate ranking, confidence calibration, and an explicit UNRESOLVABLE branch. It separates CLEAR, AMBIGUOUS, and UNRESOLVABLE states and evaluates whether calibrated uncertainty can reduce high-confidence semantic errors in social-media coreference resolution.

2 Task Definition and Data Construction

2.1 Task Definition and Label Taxonomy

Ambiguous coreference resolution is formulated as a multi-output supervised learning task. For sample i , the input is represented as:

$$X_i = (t_i, c_i, E_i, m_i) \quad (1)$$

The corresponding output is represented as:

$$Y_i = (a_i, r_i, u_i, p_i) \quad (2)$$

Eqs. (1)–(2) map the social-media snippet, thread context, candidate antecedents, and referential expression to ambiguity, ranking, UNRESOLVABLE, and calibrated candidate probabilities. CLEAR, AMBIGUOUS, and UNRESOLVABLE are mutually exclusive decision states.

2.2 Synthetic Data Construction and Verification

A controlled template pipeline covered short-video titles, comment replies, community messages, creator profiles, and anonymous feedback. It varied candidates, semantic similarity, reply depth, antecedent distance, ellipsis, and evidence availability, assigning CLEAR, AMBIGUOUS, or UNRESOLVABLE labels. AMBIGUOUS samples retained plausible candidates without decisive evidence, whereas UNRESOLVABLE lacked supported antecedents. All 86,400 samples passed checks; 12,960 (15%) underwent manual validation by three annotators assessing fluency, candidate

validity, and label correctness. Fleiss’ κ reached 0.86, with disagreements resolved through voting and senior review.

2.3 Dataset Statistics

The dataset covers five social media scenarios: short-video titles emphasize contextual ellipsis; comments and replies

emphasize cross-level reply shifts; community messages emphasize multiple-entity co-occurrence; creator profiles emphasize mixed identity references; and anonymous feedback emphasizes missing antecedents in user-generated content across online platforms. The statistical results are shown in Table 1.

Table 1: Dataset statistics

Scenario Type	Text Snippets/Entries	Entity Mentions/Count	Referential Expressions/Count	Key Constructs
Short Video Titles	17,280	36,200	12,600	Lack of Context, Vague Expressions
Comment Replies	25,920	75,600	24,800	Cross-Level Replies, Chained Navigation
Community Messages	17,280	44,800	13,400	Co-Occurrence of Multiple Entities, Candidate Competition
Creator Profiles	12,960	32,400	7,600	Identity References, Role Blending
Anonymous Feedback	12,960	25,700	8,900	Missing or Unresolvable Antecedents
Total	86,400	214,700	67,300	Multi-Scenario Joint Modeling

3 AmbiGraph-Coref Model Design

3.1 Construction of Heterogeneous Relational Graphs

AmbiGraph-Coref builds a heterogeneous graph of referential expressions, candidates, comments, replies, and contextual evidence, with edges encoding candidate, context, reply, and co-occurrence relations.

The graph structure is defined as follows:

$$G = (V, E, R), \quad V = V^m \cup V^c \cup V^e \cup V^r \cup V^s \quad (3)$$

Candidate mentions are deduplicated, filtered by type compatibility, reply accessibility, and relevance, then ranked by similarity and distance; DeBERTa-v3-base supplies 256-token node representations.

3.2 Modeling the Context of Comment Chains

Comment-chain context follows “Original Post — Direct Reply — Same-Level Comment — History Fragment,” with cross-level evidence propagated under structural-distance constraints.

$$\tilde{h}_i = \sigma(W_0 h_i + \sum_{r \in R} \sum_{j \in N_i^r} \alpha_{ij}^r W_r h_j) \quad (4)$$

Equation (4) updates each node from its own representation and relation-specific neighbors.

The attention weight combines semantic relevance with comment-chain distance as follows:

$$s_{ij}^r = \frac{(W_q h_i)^T (W_k h_j)}{\sqrt{d}}, \quad \alpha_{ij}^r = \frac{\exp(s_{ij}^r - \lambda d_{ij})}{\sum_{k \in N_i^r} \exp(s_{ik}^r - \lambda d_{ik})} \quad (5)$$

Equation (5) combines semantic relevance with distance decay; two graph layers use 768 hidden units, 0.1 dropout, and a four-level context limit.

3.3 Candidate Entity Ranking Mechanism

Candidate antecedents are ranked rather than treating ambiguous coreference as a single-label classification problem. For referential expression m and candidate entity e_i , the ranking score is computed from semantic matching, graph-relational relevance, and contextual-evidence strength:

$$s(e_i, m) = w_1 f_{\text{sem}}(e_i, m) + w_2 f_{\text{rel}}(e_i, G) + w_3 f_{\text{ctx}}(e_i, C) \quad (6)$$

where $s(e_i, m)$ is the candidate score; f_{sem} is the bidirectional semantic-matching score between the referential expression and candidate entity; f_{rel} is the heterogeneous-graph propagation score; f_{ctx} is the contextual-evidence strength; and w_1, w_2, w_3 are learnable weights. Candidate logits are temperature-calibrated to produce the Top-k probability distribution. The ambiguity trigger is then computed as the calibrated probability gap between the Top-1 and Top-2 candidates. Threshold τ is selected on the validation set by grid search from 0.05 to 0.20 in increments of 0.05 after temperature T has been fixed. Thus, both threshold selection and inference use the same calibrated distribution, avoiding inconsistent decisions caused by applying τ to uncalibrated scores before probability calibration.

4 Confidence Calibration and Error Suppression Methods

4.1 Multi-Task Joint Learning Framework

As shown in Figure 1, Shared graph representations feed ambiguity, candidate-ranking, and UNRESOLVABLE heads, allowing candidate competition and insufficient evidence to be modeled separately while retaining one shared encoder.

$$L = \sum_{t \in \{\text{amb, rank, unres, cal}\}} \exp(-\gamma_t) L_t + \gamma_t \quad (7)$$

Equation (7) jointly optimizes ambiguity, ranking, rejection, and calibration losses. Training uses batch size 32, AdamW at 2×10^{-5} , warmup ratio 0.1, and dropout 0.1.

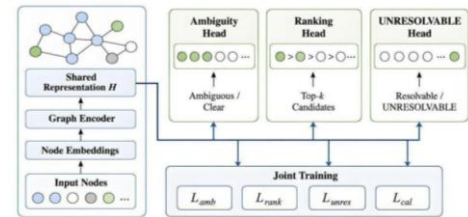


Figure 1: Multi-task learning framework of AmbiGraph-Coref

4.2 Probability Calibration of Candidate Antecedents

Temperature scaling is fitted on the validation set to prevent small ranking gaps from producing overconfident outputs.

$$p_i^{\text{cal}} = \frac{\exp(z_i/T)}{\sum_{j=1}^n \exp(z_j/T)} \quad (8)$$

Temperature is selected by validation NLL; ECE uses 15 bins.

$$\text{ECE} = \sum_{b=1}^B \frac{|B_b|}{N} |\text{acc}(B_b) - \text{conf}(B_b)| \quad (9)$$

Joint calibration lowers ECE to 0.057.

4.3 Identification of UNRESOLVABLE Samples

UNRESOLVABLE fuses learned rejection with candidate-confidence and contextual-evidence signals.

$$\begin{aligned} g_i &= \mathbb{I}[\max(p^{\text{cal}}) < \tau_1 \vee (\Delta p < \tau_2 \wedge q_{\text{ctx}} < \tau_3)], \\ p_i^{\text{hyb}} &= \beta p_i^{\text{head}} + (1 - \beta) g_i, \\ y_{\text{unres}} &= \mathbb{I}[p_i^{\text{hyb}} \geq \tau_0] \end{aligned} \quad (10)$$

Validation-tuned evidence thresholds provide a low-confidence safeguard. The parameters β , τ_0 , τ_1 , τ_2 , and τ_3 are selected on the validation set. The learned head therefore provides the primary prediction, while the evidence rule serves as a low-confidence safeguard rather than an independent decision rule.

4.4 Analysis of Semantic Error Types

Semantic errors were classified as candidate competition, comment-chain jumps, missing contextual evidence, and informal expressions. Candidate competition arose when co-

occurrence propagation made entity representations overly similar. Comment-chain jumps linked expressions to adjacent comments because of weak reply-edge weighting or excessive distance decay. Missing evidence activated the UNRESOLVABLE branch for short titles and anonymous comments. Informal abbreviations, nicknames, and omitted subjects also caused errors. Diagnostic samples were used to adjust edge weights, temperature scaling, rejection thresholds, and low-confidence resampling during training.

5 Experimental Results and Application Validation

5.1 Experimental Setup and Evaluation Metrics

The 86,400 snippets were split 7:1:2 for training, validation, and testing while preserving label distributions. Models used AdamW, a 2×10^{-5} learning rate, batch size 32, sequence length 256, and early stopping. Baselines included End-to-End Coref, BERT-Classifier, Span-Ranking, Graph-Coref, and zero-/five-shot Llama-3.1-8B-Instruct. F1, accuracy, MRR, and risk-classification metrics were reported.

5.2 Ambiguous Coreference Identification Performance

To evaluate the effectiveness of the heterogeneous relational graph for ambiguous coreference resolution, AmbiGraph-Coref was compared with multiple baseline models on the same test set, as reported in Table 2. End-to-end coreference models perform well on straightforward samples; however, they struggle with comment-chain jumps and multiple competing antecedents. BERT classifiers lack explicit constraints for candidate ranking and therefore produce high-confidence errors in cross-level replies, whereas the Span-Ranking model improves antecedent ordering but does not explicitly represent structural relations among comments and replies in user-generated social media threads.

Table 2: Comparison of ambiguous coreference identification and candidate-ranking performance

Model	Precision	Recall	F1	Accuracy	MRR
End-to-End Coref	0.742	0.701	0.721	0.736	0.781
BERT-Classifier	0.768	0.724	0.745	0.758	0.796
Span-Ranking Coref	0.801	0.769	0.785	0.793	0.823
LLM Zero-shot	0.772	0.748	0.760	0.776	0.811
LLM Few-shot	0.808	0.781	0.794	0.807	0.835
Graph-Coref without calibration	0.826	0.811	0.818	0.821	0.861
AmbiGraph-Coref	0.864	0.848	0.856	0.842	0.887

5.3 Effect of Confidence Calibration

Table 3 and Figure 3 summarize calibration behavior. Uncalibrated softmax produces the largest ECE because small Top-1/Top-2 score gaps can still yield excessive confidence on ambiguous samples. Temperature scaling smooths the

candidate distribution, while joint calibration further incorporates UNRESOLVABLE evidence to improve low-evidence decisions.

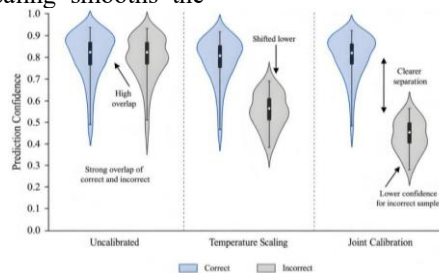


Figure 3: Confidence distributions before and after calibration

Table 3: Comparison of calibration methods

Calibration Method	ECE	NLL	Brier Score	High-Confidence Error Rate (%)	F1 Score for UNRESOLVABLE Classification
Uncalibrated softmax	0.184	0.612	0.219	18.7	0.736
Label smoothing	0.132	0.574	0.196	14.9	0.754
Temperature scaling	0.081	0.521	0.171	10.6	0.781
Temperature scaling + candidate-discrimination threshold	0.066	0.498	0.158	8.4	0.804
Joint calibration	0.057	0.472	0.149	6.9	0.821

Joint calibration reduces ECE from 0.184 to 0.057 and the high-confidence error rate from 18.7% to 6.9%. Figure 4 shows that incorrect predictions shift toward lower confidence after calibration, while correct samples retain higher-confidence density. The separation is most evident when candidate margins are small or contextual evidence is weak. Calibration therefore corrects overconfident ambiguous predictions rather than uniformly suppressing confidence across all samples, especially when contextual evidence is weak or structurally distant. This behavior is consistent with selective correction of uncertain candidate assignments.

5.4 Validation Analysis for Downstream Tasks

Downstream validation supplies calibrated candidate probabilities, ambiguity labels, and UNRESOLVABLE labels to sentiment and comment-risk classifiers while holding the encoder and training settings constant. Candidate representations are probability-weighted before task-specific prediction, and Table 4 reports the resulting classification and entity-binding measures.

Table 4: Performance comparison on downstream tasks

Input Setting	Sentiment Classification Macro-F1	Risk Identification AUC	Incorrect Entity Binding Rate (%)	High-Risk False Positive Rate (%)
Raw text only	0.713	0.764	21.6	16.8
Traditional coreference-resolution output	0.736	0.781	18.9	15.2
Top-1 candidate output	0.752	0.796	16.4	13.7
Candidate probability distribution	0.768	0.811	13.2	11.5
Ambiguity identification + UNRESOLVABLE decision	0.781	0.823	9.7	9.8

Reference-uncertainty features raise sentiment Macro-F1 from 0.713 to 0.781 and risk-identification AUC from 0.764 to 0.823; incorrect entity binding falls from 21.6% to 9.7% and the high-risk false-positive rate from 16.8% to 9.8%. Calibrated candidate distributions distinguish definite references from competing antecedents, while UNRESOLVABLE suppresses unsupported assignments when evidence is insufficient. These gains support the proposed mechanism within the synthetic evaluation setting, but the rule-derived downstream labels are not independently annotated real-platform data. Real-world generalization therefore remains to be verified on authentic social-media interactions.

6 Conclusion

A controlled dataset covering five social-media scenarios was constructed to examine ambiguous and unsupported coreference. AmbiGraph-Coref integrates heterogeneous relational propagation, candidate ranking, multitask learning, probability calibration, and an UNRESOLVABLE branch. It achieves an F1 score of 0.856, accuracy of 0.842, and MRR of 0.887. Joint calibration reduces ECE from 0.184 to 0.057 and high-confidence errors from 18.7% to 6.9%, while improving downstream sentiment classification and risk identification. However, the primarily synthetic data and labels cannot establish real-platform effectiveness. Future work should use anonymized manually annotated interactions, evaluate cross-topic and cross-platform transfer, address sarcasm and longer conversations, and simplify online inference.

REFERENCES

- [1] Bohnet B, Alberti C, Collins M. Coreference resolution through a seq2seq transition-based system[J]. Transactions of the Association for Computational Linguistics, 2023, 11: 212–226.
- [2] Baruah S, Narayanan S. Character coreference resolution in movie screenplays[C]//Findings of the Association for Computational Linguistics: ACL 2023. 2023: 10300–10313.
- [3] Žabokrtský Z, Konopík M, Nedoluzhko A, et al. Findings of the second shared task on multilingual coreference resolution[C]//Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023: 1–18.
- [4] Skachkova N, Anikina T, Mokhova A. Multilingual Coreference Resolution: Adapt and Generate[C]//Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023: 19–33.
- [5] Arslan T P, Acar K, Eryiğit G. Neural end-to-end coreference resolution using morphological information[C]//Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution. 2023: 34–40.
- [6] Novák M, Dohnalová B, Konopík M, et al. Findings of the Third Shared Task on Multilingual Coreference Resolution[C]//Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora, and Coreference. 2024: 78–96.
- [7] Pražák O, Konopík M. End-to-end Multilingual Coreference Resolution with Headword Mention Representation[C]//Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora, and Coreference. 2024: 107–113.
- [8] Semínck O, Bourgois A, Dupont Y, et al. GLaRef@ CRAC2025: Should We Transform Coreference Resolution into a Text Generation Task? [C]//Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora, and Coreference. 2025: 119–129.
- [9] Hejman J, Pražák O, Konopík M. Fine-Tuned Llama for Multilingual Text-to-Text Coreference Resolution[C]//Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora, and Coreference. 2025: 140–148.
- [10] Borszuckovszki M, De Jong I P, Valdenegro-Toro M. Know what you do not know: Robustness of verbalized uncertainty estimation on corrupted images in vision-language models[C]//Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025). 2025: 247–26