EmbiText: Embracing Ambiguity by Annotation, Recognition and Generation of Pronominal Reference with Event-Entity Ambiguity

Amna
IT University of Copenhagen
amnasheikh1@gmail.com

Christian Hardmeier
IT University of Copenhagen
chrha@itu.dk

Abstract

Consider the example "*The bird sang the nursery rhyme beautifully. It made everyone in the room smile*". The pronoun 'it' here refers either to the bird or to the event of singing. This example is inherently ambiguous. It cannot be meaningfully disambiguated as an event or entity reference, as both readings result in the same text meaning. This study introduces a new dataset **EMBITEXT** to preserve ambiguity in the language by navigating through the ambiguity surrounding the pronominal reference to the entity or event. Oftentimes, ambiguity does not necessarily need to be resolved but is modelled carefully. Furthermore, this study explores the capacity of LLMs (Llama, Mistral, Gemini, Claude AI) to embrace ambiguity in generating text that exhibit referential ambiguity via an In-Context learning approach. To evaluate of the dataset, RoBERTa was fine-tuned on this data to model ambiguity while simultaneously distinguishing between entity or event references. Results demonstrate EmbiText's capacity to advance the ongoing NLP research by modelling linguistic ambiguity in computational environments instead of fully disambiguating it, thereby retaining diverse interpretations where resolution may alter meaning.

1 Introduction

Ambiguity in language represents multiple plausible meanings, interpretations, and contexts of words, phrases, or sentences. The occurrence of ambiguity in language is inherently natural and can be a stylistic choice or a result of poetic expression, but also occur unintentionally. Humans tend to navigate around these ambiguities naturally at most times as compared to computers and machines, although in extreme cases ambiguity may lead to profound confusion for even humans. Researchers and engineers are leveraging Artificial Intelligence (AI) to navigate through this ambiguity along with its contextual understanding with a level of naturalness akin to human understanding of language.

A simple pronoun could refer to any entity, an event, or in some cases may refer to both. An entity is typically a noun denoting an object inside the discourse realm, whereas an event is a verbal phrase describing an action that has occurred. Anaphora resolution research has traditionally focused on complete disambiguation of pronominal references, as seen in corpora like OntoNotes by Weischedel et al. (2010) and GUM Corpus by Zeldes et al. (2025), yet some ambiguities are difficult to resolve or resolving them lead to discrepancies in coreference annotation. Lapshinova-Koltunski et al. (2019) This calls for a dataset, specifically curated and robust to include ambiguous examples and their potential antecedents enabling the NLP models to both detect ambiguity and quantify uncertainty.

Examples of ambiguous cases this study primarily focuses on:

- *The bird mimicked the nursery rhymes beautifully. It made everyone in the room smile.*
- *The volcano erupted violently. It created a huge crater.*
- *The garden was blooming with flowers, which made me feel refreshed.*
- *The fireworks display wonderfully lit up the sky on time. This added colours to the ceremony.*

This study underscores the linguistic fundamentals of ambiguity, reflecting on semantics, linguistic theories, and syntax to interpret how multiple readings can be extracted from pronoun reference.

This study addresses these research questions:

- How can the ambiguity inherent in pronominal references between entities and event be identified, annotated, and modeled in Natural Language?

Proceedings of the 6th Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (CODI 2025), pages 157–165

- Are LLMs capable of embracing ambiguity in natural language rather than resolving it?

We answer these questions by developing a curated dataset that models the ambiguity in pronominal references between entities and events. Data is annotated and then evaluated by fine-tuning an LLM. The text examples exhibit ambiguity surrounding pronoun reference. Additionally, LLMs are prompted to generate ambiguous examples and quantify the likelihood of pronouns referring to entities and events. This study aims to contribute to the literature in Natural Language Processing by exploring the complexities of natural language and leveraging AI to preserve linguistic ambiguity.

2 Background

The literature review explores types of ambiguities in natural language, i.e. syntactic, discourse, anaphoric, semantic, and lexical ambiguities such as in a study by Anjali and Anto (2014). Considerable focus has been on investigating ambiguities in different settings. Duzi (2013) claims that ambiguity is not only prominent in informal conversations but also evidently exists in formal discussions and arguments, particularly focusing on philosophical approaches. Chukwu (2015) discovers and resolves ambiguities and incorporates admissible ambiguity into literary writing to understand word order and context.

Researchers have developed corpora for coreference resolution. Yuan et al. (2023) introduced a corpus of sentence pairs with ambiguous and unambiguous referents to compare human and model sensitivity to ambiguity, focusing primarily on disambiguation. In contrast, EmbiText models graded ambiguity, thereby prioritizing preservation over resolution of ambiguity. Datasets like LitBank by Bamman et al. (2020), LegalCore by Wei et al. (2025), and PreCO by Chen et al. (2018) are aimed at entity-level coreference and yield efficient error analysis, but ignore combined representation of entities and events. Emami et al. (2019) introduces a context-driven coreference corpus by eliminating gender and number cues. KnowRef focuses on disambiguation, unlike EmbiText, which embraces ambiguity.

Loáiciga et al. (2017) investigates the ambiguous nature of the pronoun "it" by applying the maximum Entropy classifier to differentiate between anaphoric, event-referential, and pleonastic uses of "it", with a focus on a single type of pronoun "it" and silver-standard data. Loáiciga et al. (2020) subsequently introduced cross-lingual signals-related disambiguation system for event-based ambiguities with exclusive focus on the English pronoun 'it' and reliance on silver standard data. This narrows the scope and limits applicability to a wider range of contexts. Bevacqua et al. (2021) investigated linguistic patterns in event-entity coreference across five languages using the story continuation task, while focusing on disambiguation using a psycholinguistic approach rather than representing ambiguity as linguistic characteristic with computational models. Joshi et al. (2019) proposed efficient BERT-based system for entity coreference which struggled with encoding relations between entities. Le et al. (2022) proposed extremely accurate scientific coreference resolution with In Context learning, but is restricted by prompt-based capacity and cross-domain generalization. These studies underscore the need for extensive research, which our study aims to accomplish by curating EmbiText, annotated data to navigate through ambiguity in pronominal reference and leveraging LLMs to provide insights about the linguistic phenomena related to pronoun reference. Unlike traditional disambiguation, this study coherently embraces ambiguity by introducing data with inherently ambiguous cases.

3 Methodology

This section outlines the complete study pipeline, from data acquisition and processing to model training. Provided that the focus of the study is on pronoun reference, the selected pronouns are: 'It', 'This', 'That and 'Which'.

3.1 Data Acquisition

Georgetown University Multilayer Corpus (GUM) by Zeldes et al. (2025) was chosen for experimental analysis as it contains real-world examples from various domains such as academic, art, literature, interviews, etc. Datasets from the coreference section of GUM were selected due to their relevance to the focus of this study, containing a total of 14158 text examples. About 4860 text examples containing the selected pronouns were extracted. Each example was subjected to extensive auditing to examine whether the pronoun referred to an antecedent and to check for potential ambiguity. Examples containing dummy pronouns e.g. "*Basing letters on objects (pictographs) is an easy way to start a writing system. Try this with a group of friends. It's*

much more fun when there are other people that can understand your language" and informal dialogues i.e. "*But, but I remember, like I went there with this person, it's kind of funny*" were excluded. The '*it*' here functions as a syntactic placeholder without an antecedent. In total, 249 examples were sampled from the corpus, containing a mix of ambiguous and unambiguous examples. We conducted a text generation experiment by using Decoder-only LLMs with an In-Context Few Shot learning approach to assess their capability in generating examples containing ambiguous pronoun references. Mistral AI (`Mistral-7B-Instruct-v0.1`) by (Jiang et al., 2023), Gemini 2.0 by (Google DeepMind, 2023), Llama 3.0 (`llama-3-8b-instruct`) by (Grattafiori et al., 2024) and Anthropic's Claude AI (`Claude 3 Haiku`) (Anthropic, 2024) were selected, fine-tuned to generate texts identical to the requirement of this research. The hyperparameters setting included: `temperature` set to 0.7, `do_sample` set to true, and `max_new_tokens` set to 700 for Mistral and 1000 for Claude. This setting ensured a controlled level of linguistic creativity, diversity, and randomness in the generated examples while adhering to the task-specific prompts. With extensive prompt engineering (see 7), 120 examples were generated; 30 examples from each model.

The data examples generated from LLMs (Gemini, Mistral AI, Claude, and Llama) along with the shortlisted data examples from GUM corpus were integrated into a composite dataset for annotations.

3.1.1 Annotations

To identify ambiguity surrounding pronominal references to entities or events, text examples were annotated. Dataset was annotated using Label Studio by Tkachenko et al. (2025). Each text example was critically examined and classified as ambiguous when it contained both entity and event references or as unambiguous when it contained only one reference type. To ensure an unbiased and systematic annotation process, both authors of this study independently annotated the examples. This strategy was used to mitigate individual bias. The custom labeling setup involved rating examples on an 11-point scale ranging from 100% entity-leaning to 100% event-leaning. Annotators labeled and rated each example: Figure 1 illustrates the star icon used for annotations. Annotators used their contextual understanding and linguistic understanding while following annotation guidelines.

We initially calculated the inter-annotator agreement while correcting for chance agreement by using Cohen's Kappa by Cohen (1960) and ordinal Krippendorff's alpha by krippendorff (2004). This was followed by annotators adopting an adjudication approach. Both annotators jointly reviewed all cases of annotation disagreements and resolved them through systematic discussion, resulting in consensus for each case. This process led to a revised annotated dataset, devoid of inter-annotator disagreements and was subsequently used as final training and test data for the model.

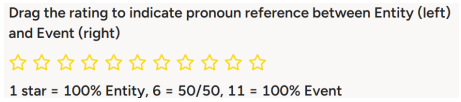


Figure 1: Star rating icon; the first four stars starting from the left represent pronoun reference leaning towards entity while the first star represents 100% entity reference, the three stars in the middle denote ambiguity, the last four stars represent pronoun reference leaning towards event, with the last and eleventh star indicating 100% event.

An example of an ambiguous case is as follows: "*The dog barked at the mailman. It startled the children*". Here, either the dog (entity) or 'the barking of dog' (event), startled the children, hence entity and event probability both receive values of 0.5. Probabilities are categorized for a simple annotation process; 0.1-0.39 represents entity-leaning while remaining 0.69-0.9 represent event-leaning and vice-versa, and 0.4-0.6 represent Ambiguous cases. These five labels were later condensed into three labels: entity leaning, event-leaning and ambiguous by removing 'entity' and 'event' labels. Three-label scheme categorizes probabilities in ranges of: 0-0.39 for entity-leaning and the remaining 0.69-1 for event-leaning. Refer to appendix 7 for an example. This approach introduced simplicity in labeling examples and subsequently helped mediate inter-annotator disagreement. In view of computation, merging exact entity-event categories into entity-event-leaning categories reduces sparsity. An overview of example categories in dataset is illustrated in figure 1.

Category	Number of Text Examples
Entity Leaning	127
Ambiguous	69
Event Leaning	53
Total	249

Table 1: Distribution of annotated examples across categories in the EmbiText.

3.1.2 Model Training

Transformer-based RoBERTa by Liu et al. (2019), was fine-tuned on EmbiText to test its interpretability when the pronominal reference potentially leads to multiple interpretations. The model then predicts the probability of the entity or an event, and the complementary probability is calculated as 1 minus the predicted probability, e.g. $p(\text{entity}) = 1 - p(\text{entity})$ and vice versa. For instance, if the entity prediction is 0.72, the corresponding event prediction is 0.28.

We used the HuggingFace Transformers framework to fine-tune the model. The best model checkpoint was chosen on the basis of the validation loss. Appendices 7 and 7 show the input features and hyperparameter configuration. This hyperparameter configuration is widely used as the RoBERTa fine-tuning setting for small datasets, e.g. Wolf et al. (2020), resulting in stable convergence without overfitting.

The model uses sigmoid activation function to provide a probabilistic illustration of whether a pronoun refers to an entity or an event. Probability tokens are the targets for the model to compute loss using Mean Squared error (MSE) to project the difference between predictions and ground-truth probabilities by penalizing large deviations to train the model on necessary fine-grained contextual cues. Probabilistic outcomes enable the model to project ambiguity and degrees of entity/event-leaning instead of predicting binary choices. The output was post-processed using a threshold function to map probabilities to categories of: 1) Entity-leaning, 2) Ambiguous, 3) Event-leaning.

To evaluate the system's performance, we compared it with an instruction-tuned baseline language model, Flan-T5, encoder-decoder-based system by Chung et al. (2022). It has shown promising performance across zero- and few-shot prompts setup. We conducted few-shot prompting by including a random sample of training examples. The model was responsible for generating output in the form of probability estimates for pronominal references to entities or events. The generated probabilities were categorized using the same thresholding method used to fine-tune RoBERTa.

4 Results

Our results demonstrate that EmbiText effectively embraces pronominal ambiguity, supporting its relevance for human-computer interactions. The overall results highlight the reasonable quality of the annotations of the data for this experiment. The annotators reviewed their annotated examples and disagreed on approximately 15% of the total examples. Cohen's Kappa evaluation on the initial five-label scheme resulted in a fair agreement according to (Landis and Koch, 1977) but constrained consensus between both annotators with a value of 0.24. After reducing the labels to three, Cohen's value increased to 0.36, highlighting improvement and fair agreement (Landis and Koch, 1977). Similarly, the ordinal Krippendorff's alpha value improved from 0.31 to 0.46. Despite the improvement, the score still hints at low inter-annotator agreement, reflecting the subjectivity and difficult nature of differentiating between ambiguous and entity-event-leaning pronominal references. The initial five-label scheme in Appendix 7, shows strong agreement on "ambiguous" cases but prominent disagreement on cases between "Entity", "Entity leaning" and "Ambiguous". Contrastingly, as observed in figure 2, three-label scheme clarifies that the disagreement primarily is prominent between entity-leaning and the customized labeling setup involved rating examples on an 11-point scale ranging from 100% entity-leaning to 100% event-leaning. Annotator 1 labeled more examples as ambiguous, while Annotator 2 leaned towards entity-specific labels. See Appendix 7 for cases of inter-annotator disagreement and their resolution. This enabled the annotators to resolve the discrepancies through systematic discussions of each conflicted example until a common ground was established. Subsequently, the reconciled annotations were used as the final data to train and evaluate the model.

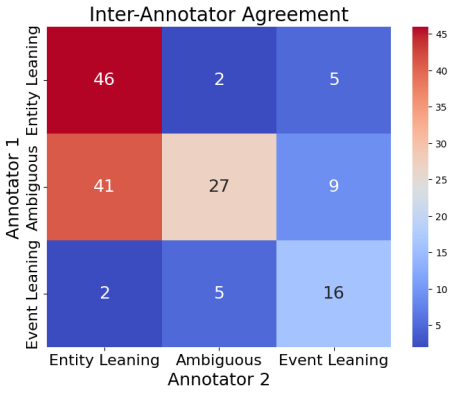


Figure 2: Confusion Matrix denoting inter-annotator agreement using three-labels scheme.

Findings from text generation experiment demonstrate that LLMs are capable of generating text ex-

amples that are ambiguous in nature. Figure 3 illustrates the superiority in performance of Llama, followed by mistral AI, however 60% of its output displayed a negative tone, emphasizing disasters, death and destruction despite the inclusion of positive and neutral examples in the prompt. This suggests that the output represents a subset of the distribution of possible examples. For example: "*The tsunami hit the shore with huge waves. It caused widespread destruction and loss of life*". Claude and Gemini demonstrated underwhelming performance.

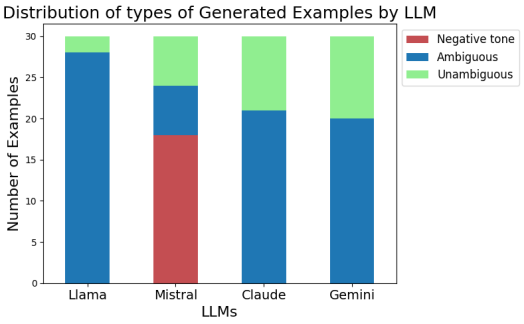


Figure 3: Overview of generated examples from LLMs.

Fine-tuning RoBERTa on this data showcased a lower rate of prediction errors: MSE of `0.019` (entity) and `0.1328` (event), RMSE of `0.3029` (entity), `0.3644` (event), and MAE of `0.2573` (entity) and `0,3119` (event), indicating more accurate results for entity references and overall reflecting a low value of average prediction error and deviation of predictions from ground truth values. Table 2 presents the comparison of our system (Fine-tunes RoBERTa) with baseline (Flan-T5). Our model yielded a lower error rates (MSE and RMSE and more accurate performance as compared to Flan-T5. This detail suggests that model succeeded in predicting probabilities values closer to true probability values.

Metric	Flan-T5 (few-shot)	RoBERTa (fine-tuned)
MSE	0.1063	**0.0847**
RMSE	0.3260	**0.2911**
Accuracy	0.314	**0.353**
Macro F1	0.096	**0.205**

Table 2: Comparison of the baseline (Flan-T5, few-shot) and our system (fine-tuned RoBERTa) on the test set, predicting entity probabilities.

5 Discussion and Conclusion

Our results corroborate that the curated EmbiText dataset and fine-tuned LLMs efficiently model natural ambiguity, especially in cases in common language where resolution is challenging. The proposed dataset demonstrate linguistic relevance and careful annotation approaches with systematic reconciliation of inter-annotator disagreements to mitigate bias and subjectivity. During annotation, some ambiguous examples featured event spans that elaborated on entities rather than individual actions i.e. "*Tomorrow, when this image is shared with the world,it will be a historic moment for science and technology*". The baseline comparison revealed that fine-tuned RoBERTa produced an improved probability calibration with a balanced distribution of categories, while Flan-T5, despite strong predictions, reflected a slight bias toward *Ambiguous* category.

Data	it	that	which	this
LLM-generated examples	108	0	0	0
GUM Corpus	28	13	7	11
Total	136	13	7	11

Table 3: Counts of selected pronouns across data examples.

The examples resulting from text generation experiments reflect the ambiguity found in natural language, where pronouns can refer to multiple antecedents and visualize multiple contextual interpretations. The results from the fine-tuned RoBERTa configuration suggest that the curated dataset accommodates referential ambiguity while distinguishing between entity and event references rather than resolving it. LLM generated text only included the pronoun "it", despite the prompts including other pronouns, suggesting further prompt refinement as seen in Figure 3. This demonstrates the ability of modern AI systems to interpret syntactic and semantic ambiguity in ways that project human-like sensitivity to multiple contexts through prompt engineering and fine-tuning. Although LLM performance is not equivalent to human cognition, the results support the second goal of this study: modeling ambiguity as a linguistic feature rather than resolving it. This study focuses on embracing ambiguity as a feature that uncovers deeper textual interpretations rather than a flaw, something previous research had neglected. This study contributes to applications involving human-computer interaction, i.e. customer service bots, dialogue systems, and assistive technologies.

6 Limitations

Our focus is primarily on the ambiguity arising from the pronominal reference between entity-event in English-specific text examples. Despite a fair agreement value and consensus-based resolution of disagreements, perception of ambiguity remains subjective, reflecting the level of difficulty of this task. Limited data size and label imbalance can cause differences between entity and event results. Additionally, the baseline Flan-T5 model is an instruction-tuned sequence-to-sequence model which makes probability prediction and classification tasks less direct as compared with encoder.only RoBERTa. Future direction should expand data size to enhance model generalizability, apply resampling techniques (i.e. SMOTE), involve cross-lingual analysis, enhanced prompt engineering techniques for text generation, employ other pretrained models as baseline models to compare with, multiple annotators (4-6) and multiple evaluators to evaluate generated examples for robust assessment and test different model architectures.

References

M. K. Anjali and P. Babu Anto. 2014. Ambiguities in natural language processing. *International Journal of Innovative Research in Computer and Communication Engineering*, 2:392–394. Accessed: 2025-05-05.

Anthropic. 2024. Claude 3 haiku (version: claude-3-haiku-20240307). Large language model developed by Anthropic.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. *Preprint*, arXiv:1912.01140. Accessed: 2025-07-20.

Luca Bevacqua, Sharid Loáiciga, Hannah Rohde, and Christian Hardmeier. 2021. Event and entity coreference across five languages: Effects of context and referring expression. *Dialogue & Discourse*, 12(2):192–226. Accessed: 2025-09-15.

Hong Chen, Zhenhua Fan, Hao Lu, Alan L. Yuille, and Shu Rong. 2018. Preco: A large-scale dataset in preschool vocabulary for coreference resolution. *Preprint*, arXiv:1810.09807. Accessed: 2025-07-20.

Ephraim Chukwu. 2015. Understanding linguistic ambiguities for the effective use of english. *Awka Journal of English Language and Literary Studies*, 6. Accessed: 2025-05-09.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416. Accessed: 2025-09-18.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46. Accessed: 2025-09-18.

Marie Duzi. 2013. *Ambiguities in natural language and ontological proofs.*, pages 179–218. Accessed: 2025-05-05.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics. Accessed: 2025-09-15.

Google DeepMind. 2023. Gemini. Accessed: 2025-04-17.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. Accessed: 2025-06-02.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825. Accessed: 2025-04-21.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Accessed: 2025-05-05.

klaus krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Departmental Papers (ASC)*, 38. Accessed: 2025-09-18.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. Accessed: 2025-09-15.

Ekaterina Lapshinova-Koltunski, Sharid Loáiciga, Christian Hardmeier, and Pauline Krielke. 2019. Cross-lingual incongruences in the annotation of coreference. In *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora*

and Coreference, pages 26–34, Minneapolis, USA. Association for Computational Linguistics. Accessed: 2025-09-18.

Nghia T. Le, Fan Bai, and Alan Ritter. 2022. Few-shot anaphora resolution in scientific protocols via mixtures of in-context experts. *Preprint*, arXiv:2210.03690. Accessed: 2025-05-10.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692. Accessed: 2025-05-25.

Sharid Loáiciga, Liane Guillou, and Christian Hardmeier. 2017. What is it? disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1325–1331, Copenhagen, Denmark. Association for Computational Linguistics. Accessed: 2025-05-28.

Sharid Loáiciga, Christian Hardmeier, and Asad Sayeed. 2020. Exploiting cross-lingual hints to discover event pronouns. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 99–103, Marseille, France. European Language Resources Association. Accessed: 2025-05-28.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Kangda Wei, Xi Shi, Jonathan Tong, Sai Ramana Reddy, Anandhavelu Natarajan, Rajiv Jain, Aparna Garimella, and Ruihong Huang. 2025. Legalcore: A dataset for event coreference resolution in legal documents. *Preprint*, arXiv:2502.12509.

Ralph Weischedel, Mitch Marcus, Martha Palmer, Eduard Hovy, Robert Belvin, Sameer Pradhan, and Lance Ramshaw. 2010. Ontonotes: A large training corpus for enhanced processing. In *Handbook of Natural Language Processing and Machine Translation*. Springer, New York, NY. Accessed: 2025-07-27.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. Accessed: 2025-09-05.

Yuewei Yuan, Chaitanya Malaviya, and Mark Yatskar. 2023. Ambicoref: Evaluating human and model sensitivity to ambiguous coreference. *Preprint*, arXiv:2302.00762. Accessed: 2025-07-20.

Amir Zeldes, Tatsuya Aoyama, Yang Janet Liu, Siyao Peng, Debopam Das, and Luke Gessler. 2025. eRST: A signaled graph theory of discourse relations and organization. *Computational Linguistics*, 51(1):23–72. Accessed: 2025-02-10.

7 Ethical Considerations

AI tools were used to assist with the polishing and enhancing the writing of this paper i.e: > - checking grammar, > - Improving and shortening text. AI assistive tools such as OpenAI's ChatGpt and Copilot were leveraged in: > - Debugging Synthetic Text generation task. > - Debugging and refining code for data preprocessing and visualization, model training loop and evaluation practices. Importantly, all experimental designs, data annotation, model evaluation and interpretations were conducted independently by the author. The disclosure of usage of AI tools is in the interests of transparency in the research process and academic integrity.

A. Input Features

Train Dataset	Test Dataset
text_example	text_example
pronoun	pronoun
entity_candidate	entity_candidate
event_candidate	event_candidate
entity_prob	
event_prob	

Table 4: Overview of train and test set features.

B. Hyperparameters

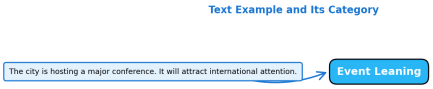
Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	16
Number of Epochs	10
Early Stopping Patience	2
Dropout Rate	0.3
Weight Decay	0.01
Gradient Clipping	0.01

Table 5: Hyperparameter configuration for RoBERTa fine-tuning.

C. Prompt for Ambiguous text examples Generation using LLMs

```
prompt = """ Generate 30 ambiguous
sentences where a pronoun could
refer to either an entity or an
event. Here are some examples:
'The volcano erupted violently.
It created a huge crater.' 'The bird
sang perfectly. It made everyone in
the room very happy.' 'Garden was
blooming with flowers, which made me
feel refreshed.' Now, generate more
examples: """
```

D. Annotation Example



E. Confusion Matrix for Five-Label scheme

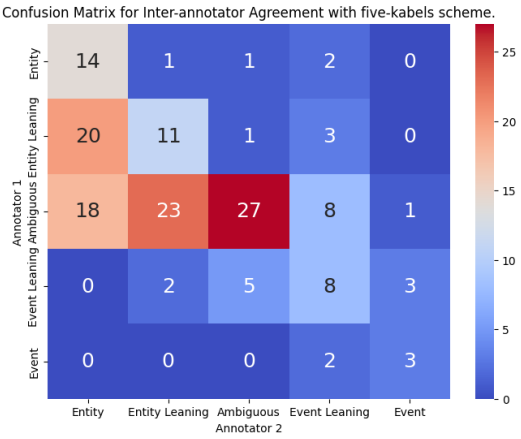


Figure 4: Confusion Matrix denoting inter-annotator agreement using five-labels scheme.

F. Annotation Disagreement and Resolution

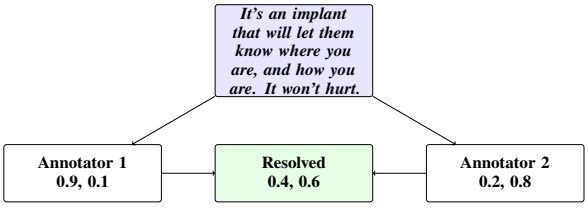


Figure 5: Annotation disagreement and resolution: unified probability distribution.

G. Annotation Disagreement and Resolution 2.0

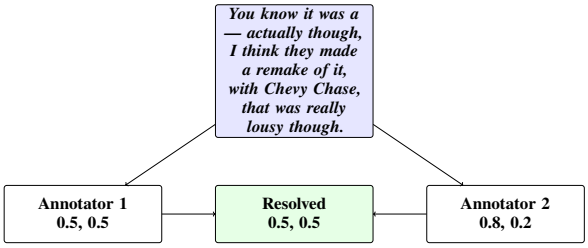
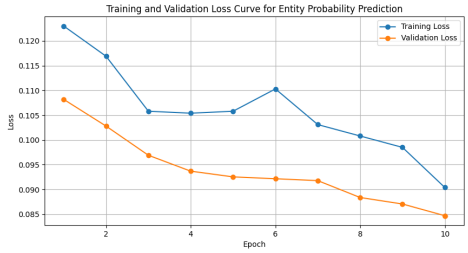
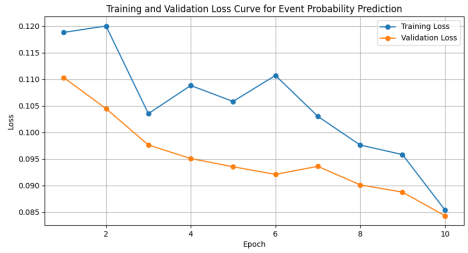


Figure 6: Annotation disagreement and resolution: unified probability distribution.

H. Loss Curve Comparison



(a) Loss curve for our model predicting entity probabilities.



(b) Loss curve for our model predicting event probabilities.

Figure 7: Training loss curves for the model predicting entities and events.

I. Metrics for Event Probabilities

Metric	Flan-T5 (few-shot)	RoBERTa (fine-tuned)
MSE	0.1063	**0.0843**
RMSE	0.3260	**0.2903**
Accuracy	0.314	**0.372**
Macro F1	0.096	**0.214**

Table 6: Comparison of the baseline (Flan-T5, few-shot) and our system (fine-tuned RoBERTa) on the test set, predicting event probabilities.