Cracks in the Bridge—or A Bridge Too Far? Comparing Human and LLM Errors in the Annotation of Bridging Anaphora

Lauren Levine and Amir Zeldes
Georgetown University
Department of Linguistics
{lel76, amir.zeldes}@georgetown.edu

Abstract

In this paper, we perform an error analysis on human and LLM annotation data from the recent GUMBridge corpus for varieties of bridging anaphora. We explore the distribution of precision and recall errors made by annotators and how that distribution correlates with bridging subtypes. We find that while LLMs perform substantially worse than human annotators, they are more balanced in their precision and recall scores than humans, whose performance strongly favors precision. With regard to subtypes, we find that COMPARISON and MERONOMY relations are easier to reliably annotate than the more broadly construed ENTITY relations for both human and LLM annotators, but that LLM errors are more distributed across subtypes than human errors. Analyzing these results, we provide insights for future annotation projects on bridging anaphora.

1 Introduction

A "bridging anaphor" is a newly introduced entity whose referent is inferable specifically due to its relation to a previously introduced entity in the discourse. Consider the following example:

(1) There is a house. **The door** is red.[1]

In the example above, the entity "**the door**" (the bridging anaphor) is specifically understood to be the door of the aforementioned "house" (the associative antecedent).

Bridging has been shown to be a difficult phenomenon to annotate, as it is dependent on the annotator's subjective understanding of entities and entity relations in a discourse (Levine and Zeldes, 2025). Analyzing the different errors made by human and LLM annotators allows us to gain insights into the strengths and limitations of each annotation approach, which can be leveraged in future efforts for the annotation of bridging anaphora, including the formulation of prompts for LLM-based approaches.

In this short paper, we investigate the following questions to better understand the errors that occur when annotating bridging:

RQ1 What types of errors (precision vs recall) dominate human and LLM annotation of bridging anaphora?

RQ2 How do sub-varieties of bridging anaphora correlate with precision and recall errors in human and LLM annotations?

2 Related Work

"Bridging" occurs when a discourse participant constructs an implicature from the entity they are currently processing back to an antecedent entity (Clark, 1975). Bridging has been studied from a variety of theoretical perspectives (e.g., Hawkins, 1978; Prince, 1981; Asher and Lascarides, 1998; Baumann and Riester, 2012), and linguistic resources have been constructed for various languages, including English (Markert et al., 2012, Rösiger, 2018, Poesio and Artstein, 2008; Uryupina et al., 2019), German (Schweitzer et al., 2018, Eckart et al., 2012), Polish (Ogrodniczuk and Zawisławska, 2016), and Czech (Nedoluzhko et al., 2009). While there have been various neural and rule-based systems for the identification of bridging instances (Rösiger et al., 2018, Yu and Poesio, 2020, Kobayashi et al., 2022), and other studies of LLMs' discourse capabilities (Li and Carenini, 2026), studies of LLMs' abilities to detect bridging anaphora are limited, with minimal benchmarks available (Bu et al., 2025).

3 Data

For our investigation of errors in the annotation of bridging anaphora, we leverage data from the

[1] Bridging anaphora are marked in bold face, and their associative antecedents are underlined.

GUMBridge corpus, a recent effort to annotate sub-varieties of bridging anaphora in English (Levine and Zeldes, 2026a). Built on top of the multi-genre GUM corpus (Zeldes, 2017), which includes pre-existing entity mention and coreference annotations (see Zeldes 2022 for detailed discussion), it uses an information status based definition of bridging, and a schema of sub-varieties with 3 main categories for bridging relations: COMPARISON relations, ENTITY relations, and SET relations. The size of the partitions of GUMBridge (in terms of tokens and gold bridging instances) are shown in Table 1 (see Levine and Zeldes 2026b for a descriptive analysis of the corpus).

In order to investigate human errors in annotation, we compare data from Levine and Zeldes 2026a's existing inter-annotator agreement (IAA) study on the annotation of the GUMBridge dev set with the final adjudicated gold version of the dev set. In the IAA study, the 32 documents in the dev set (~30k tokens) were double annotated by pairs of annotators. These annotators were provided with extensive annotation guidelines and participated in several hours of training prior to completing the annotation task. Please see Levine and Zeldes 2026a for further details on the annotation procedure.

To investigate LLM errors in annotation, we compare LLM predictions on the GUMBridge test set with the corresponding gold annotations. Levine and Zeldes 2026a provides several LLM baselines for bridging resolution on the test set, providing scores for several models. We use the judgments from the best performing model, `GPT-5`, for our error analysis. We also provide results for another SoTA LLM, `gemini-3.1-pro-preview`, using the same prompts/workflow as in Levine and Zeldes 2026a, in order to have double annotated SoTA LLM data to analyze. A brief summary of this pipeline and the prompts used is included in Appendix C.

All data analyzed in this paper is publicly available with the GUMBridge v1 release, including our additional LLM annotations.[2]

4 RQ 1: Distribution of Error Types for Human and LLM Annotators

Table 2 shows the performance of human and LLM annotators on the identification of bridging instances, i.e., the identifying the bridging anaphora **and** correctly resolving it back to its associative antecedent. Partial matches, i.e., correctly identifying a bridging anaphor but not identifying the correct associative antecedent, are counted as incorrect. Avgerage and individual scores from the double annotation of the dev set are reported, as are the average and individual performances of the 2 LLMs on the test set. As one would expect, we observe that the performance of LLMs is substantially weaker than human annotators (avg. F1-score Δ of 32.9). This aligns with recent work showing that LLMs excel at tasks which are solvable by less trained crowd workers, but often fail to reach the level of skilled annotators on complex tasks (Kasner et al., 2026). Additionally, we observe that the precision and recall scores pattern differently for humans and LLMs. Humans have notably higher precision scores than recall scores (avg. Δ of 17.3), while the precision and recall scores of the LLM baselines are much closer (avg. Δ of 0.8).

To further explore the distribution of error types, in Figure 1 we show the distribution of False Positive (FP) precision errors and False Negative (FN) recall errors for humans and LLMs, including whether one or both annotators made the same er-

GUMBridge v1	Tokens	Bridging Instances	Bridging per 1k Tokens
Train	213k	4k	18.9
Dev	30k	732	24.5
Test	30k	562	18.6
Test2	18k	379	21.2
Total	291k	5.7k	19.6

Table 1: Distribution of bridging instances across GUMBridge partitions.

	Precision	Recall	F1-Score
Humans (dev)			
Annotator 1	68.5	56.4	61.9
Annotator 2	67.2	44.8	53.8
Avg.	67.9	50.6	57.9
LLMs (test)			
GPT-5	23.2	20.3	21.7
Gemini	**27.6**	**28.9**	**28.2**
Avg.	25.4	24.6	25.0

Table 2: Performance of human and LLM annotators identifying bridging pairs (on dev and test respectively).

[2]GUMBridge data, code, and LLM results publicly available: https://github.com/lauren-lizzy-levine/gumbridge/

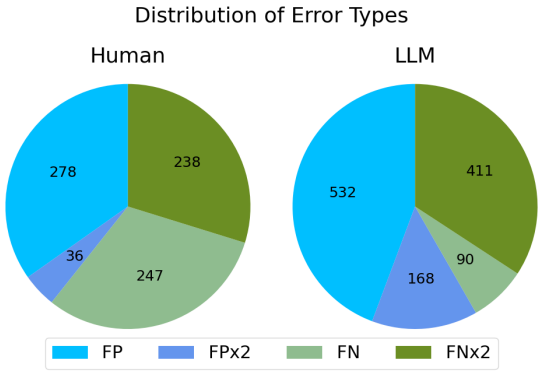

Figure 1: Distribution of error types for human and LLM annotators. FP and FN refer to errors committed by a single annotator, while FPx2 and FNx2 refer to errors committed by both annotators.

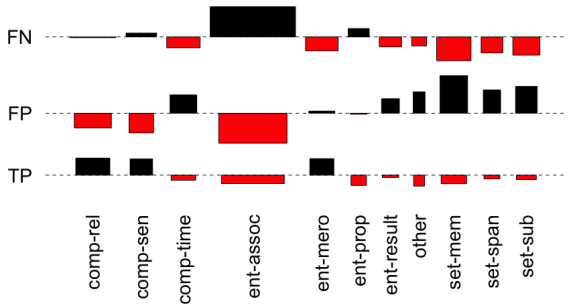

Figure 2: χ^2 residuals for human error type (False Negatives (FN), False Positives (FP), True Positives (TP)) and bridging subtype label. (X-squared = 168.73, df = 20, p-value < 2.2e-16)

ror.[3] The FP and FN categories are for instances where an error was committed by only a single annotator, while the FPx2 and FNx2 categories are for instances where the same error was committed by both annotators. We once again see that human annotators have a higher proportion of recall errors, while the LLMs now have a higher proportion of precision errors. This indicates that the two LLMs were more varied in their False Positives than their False Negatives, and that they are more likely to over-generate than their human counterparts.

Additionally, we note that the human False Negatives are roughly evenly split between being missed by one annotator or both annotators, while the LLM False Negatives are dominated by instances missed by both models. This suggests that having multiple human annotators will be more valuable than having multiple LLM annotators when trying to have broad coverage for identifying bridging anaphora, as one human annotator is more likely than an LLM to notice what another annotator has missed. However, variation in LLM errors suggests that querying a broader range of models remains valuable, as this diversity is essential for ensembling/voting approaches to bridging anaphora identification.

5 RQ 2: Analysis of Subtype Errors for Human and LLM Annotators

Every instance of bridging in the GUMBridge corpus has a subtype annotation which contains one or more subtype labels. These subtype labels fall into three main categories:

COMPARISON Relations The anaphor is preceded by a descriptor which implies a comparison to the antecedent (or vice versa) .

(2) I think her dog is nice, but I want to get **a bigger dog**.

ENTITY Relations The anaphor is an attribute or associated entity of the antecedent (or vice versa).

(3) There is a library around the corner. **The books** are fantastic.

SET Relations There is a set/subset relation between the bridging anaphor and antecedent.

(4) My niece got several toys for her birthday. Her favorite is **the doll**.

Within these three main categories, GUMBridge distinguishes 10 sub-varieties, and there is an additional OTHER category for a total of 11 sub-varieties (see Appendix A for details).

In this section, we investigate whether the subtype(s) of a bridging instance influences whether a human or LLM annotator will identify it. To do this, we examine the distribution of bridging subtype labels and observed error types (including true positives) for human and LLM annotators. Subtype labels for True Positives and False Negatives are taken from the gold labels in GUMBridge dev and test. The False Positive labels are taken from annotator judgments.[4] In Figure 2, we show the residuals from a χ^2 test for human error type and

[3] χ^2 for error type (FP, FPx2, FN, FNx2) and annotator type (human, LLM) is significant: X-squared = 212.09, df = 3, p-value < 2.2e-16

[4] False Positives are not included in the LLM analysis because the LLM pipeline was divided into subtasks, and the LLMs were not queried on subtype classification for FPs.

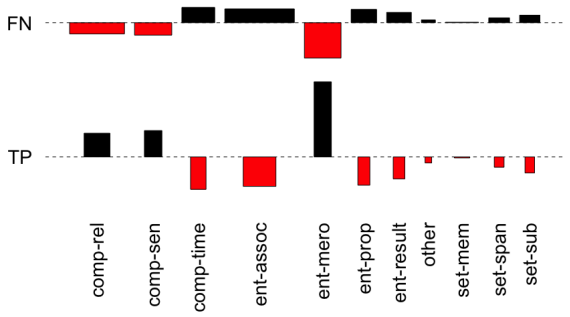

Figure 3: χ^2 residuals for LLM error type (False Negatives (FN), True Positives (TP)) and bridging subtype label. (X-squared = 87.16, df = 10, p-value = 1.955e-14)

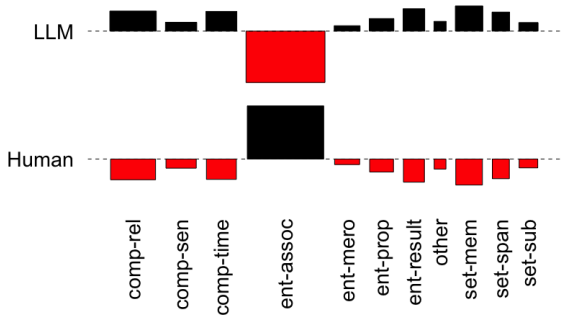

Figure 4: χ^2 residuals for human and LLM False Negatives and bridging subtype label. (X-squared = 61.472, df = 10, p-value = 1.906e-09)

bridging subtype label, and in Figure 3, we show the same for LLM error type.[5]

Looking at Figures 2 and 3, we see that both human and LLM annotators have a higher proportion of COMPARISON-RELATIVE, COMPARISON-SENSE, and ENTITY-MERONOMY in their True Positives, which indicates that these subcategories are more reliably identified by both humans and LLMs. This tendency likely reflects the fact that these subtypes frequently have overt markers which make them easier to recognize. For instance, consider the following example of COMPARISON-RELATIVE/COMPARISON-SENSE:

(5) I just had a piece of cake, and I want to have **another one**.

In the example above, "another" is a comparative marker which helps to identify the example as an instance of COMPARISON-RELATIVE, and "one" is a common lemma in instances of sense anaphora, which helps to identify the example as an instance of COMPARISON-SENSE. Both items appear in the guidelines and in LLM prompt examples. Looking at the False Negatives row, we also see that instances labeled ENTITY-ASSOCIATIVE are the more common ones to overlook for both humans and LLMs. This is unsurprising, as associative entity relations comprise the broadest sub-variety, covering a variety of implicit relations which lack overt makers, such as relational nouns (e.g., a business → **the customer**), implicit arguments (e.g., a murder → **the victim**), and prototypical associations (e.g., a wedding → **the reception**).

When we look at the human False Positives row in Figure 2, we can see that COMPARISON-TIME relations and the SET relations are more commonly mistaken for bridging by human annotators. This is likely because these semantic relations can occur without being anaphoric (i.e., they can be interpreted without reference to the antecedent) but are easily recognizable. As such, they may get flagged by annotators even if they don't fit the anaphoric criteria for bridging anaphora. Consider the following example:

(6) The 20th century was a time of rapid advancement. In particular, **the 1960s** were an eventful period.

While the above is semantically an example of a SET-SPAN-INTERVAL/COMPARISON-TIME relation, a discourse participant does not require "the 20th century" to interpret "the 1960s", which is understandable by itself (though by contrast, "the 60s" could bridge from a particular century).

While the general patterns for which subtype relations are easier are relatively consistent between LLMs and human annotators, we can also compare the two directly by looking at the subtype label distribution of the False Negative occurrences for humans and LLMs. In Figure 4, we show the residuals from a χ^2 test for human and LLM False Negatives and bridging subtype label. We see very clearly that human errors of omission are concentrated on instances of ENTITY-ASSOCIATIVE, while the errors by LLMs are more spread across subtypes. However, a plurality of missed bridging instances for the LLMs are ENTITY-ASSOCIATIVE (see Appendix B), and as Figure 3 shows ENTITY-ASSOCIATIVE is actually a difficult subtype category for LLMs,

[5]The raw counts of the subtype labels that appear in each error category are given in Appendix B

just as it is for humans. As such, it is not the case that LLMs perform very well on associative bridging instances, just that their errors are more spread across the subtypes, while humans are more strongly concentrated on ENTITY-ASSOCIATIVE.

6 Conclusion: Takeaways for Future Annotation Efforts

In this paper, we compared human and LLM errors in the annotation of bridging anaphora. Looking at the error distributions of human and LLM annotators, we saw that:

- LLM annotators are worse overall
- Humans favor precision over recall, while LLM are more balanced between the two
- Human variability can provide broader coverage of bridging than LLMs

Looking at the subtype distributions in the errors of human and LLM annotators, we saw that:

- Human and LLM annotators find the same bridging subtypes easier to identify
- Humans are prone to incorrectly identifying SET relations as bridging instances
- Relative to humans, LLM errors are more distributed across bridging subtypes

Based on these findings, we see that for the time being there is a clear advantage to using human annotators when creating data for bridging anaphora, and that human annotation projects should focus on finding a way to boost recall. In this regard, LLMs may be useful for selecting candidates for human annotators to consider as bridging anaphora. LLMs may also be useful in flagging potential human errors due to inattention or misapplication of guidelines (see Nahum et al. 2025; Chochlakis et al. 2025). Finally, future annotation guideline refinement should focus on further specifying the ENTITY-ASSOCIATIVE subtype, as it is the most difficult sub-variety for both human and LLM annotators, as well as adding attention to non-anaphoric set relations being mistaken for bridging anaphora.

Limitations

This effort focuses on performing error analysis on annotation data produced by a previous paper. As such, it is constrained by the limitations of the data produced in that work (such as LLM False Positives not having subtype annotations) and does not create new bridging data besides an additional LLM baseline. However, the comparison of human and LLM errors in the annotation of bridging anaphora is previously unexplored, and we hope the results of this analysis provide novel insights for annotation work.

References

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. *Prosody and meaning*, 25:119–162.

Lanni Bu, Lauren Levine, and Amir Zeldes. 2025. DiscoTrack: A multilingual LLM benchmark for discourse tracking. *arXiv preprint arXiv:2510.17013*.

Georgios Chochlakis, Peter Wu, Tikka Arjun Singh Bedi, Marcus Ma, Kristina Lerman, and Shrikanth Narayanan. 2025. Humans hallucinate too: Language models identify and correct subjective annotation errors with label-in-a-haystack prompts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19637–19656, Suzhou, China. Association for Computational Linguistics.

Herbert H. Clark. 1975. Bridging. In *Theoretical Issues in Natural Language Processing*.

Kerstin Eckart, Arndt Riester, and Katrin Schweitzer. 2012. A discourse information radio news database for linguistic analysis. In *Linked Data in Linguistics: Representing and Connecting Language Data and Language Metadata*, pages 65–76, Berlin, Heidelberg. Springer Berlin Heidelberg.

John A. Hawkins. 1978. Definiteness and indefiniteness: A study in reference and grammaticality prediction. *Journal of Linguistics*, 27:405–442.

Zdeněk Kasner, Vilém Zouhar, Patrícia Schmidtová, Ivan Kartáč, Kristýna Onderková, Ondrej Platek, Dimitra Gkatzia, Saad Mahamood, Ondrej Dusek, and Simone Balloccu. 2026. LLMs as span annotators: A comparative study of LLMs and humans. In *Proceedings of the First Workshop on Multilingual Multicultural Evaluation*, pages 1–22, Rabat, Morocco. Association for Computational Linguistics.

Hideo Kobayashi, Yufang Hou, and Vincent Ng. 2022. Constrained multi-task learning for bridging resolution. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 759–770, Dublin, Ireland. Association for Computational Linguistics.

Lauren Levine and Amir Zeldes. 2025. Subjectivity in the annotation of bridging anaphora. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 48–59, Vienna, Austria. Association for Computational Linguistics.

Lauren Levine and Amir Zeldes. 2026a. Gumbridge: A corpus for varieties of bridging anaphora. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 6823–6837, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

Lauren Levine and Amir Zeldes. 2026b. What's in a bridge?: A descriptive, multi-genre analysis of the gumbridge corpus for varieties of bridging anaphora. In *Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences (7th CODI) and Computational Models of Reference, Anaphora and Coreference (9th CRAC)*, San Diego, California, USA. Association for Computational Linguistics.

Chuyuan Li and Giuseppe Carenini. 2026. BeDiscovER: The benchmark of discourse understanding in the era of reasoning language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4417–4479, Rabat, Morocco. Association for Computational Linguistics.

Katja Markert, Yufang Hou, and Michael Strube. 2012. Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 795–804, Jeju Island, Korea. Association for Computational Linguistics.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26782–26809, Suzhou, China. Association for Computational Linguistics.

Anna Nedoluzhko, Jiří Mírovský, and Petr Pajas. 2009. The coding scheme for annotating extended nominal coreference and bridging anaphora in the Prague dependency treebank. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 108–111, Suntec, Singapore. Association for Computational Linguistics.

Maciej Ogrodniczuk and Magdalena Zawisławska. 2016. Bridging relations in Polish: Adaptation of existing typologies. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, pages 16–22, San Diego, California. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ellen F. Prince. 1981. Toward a taxonomy of given-new information. *Radical pragmatics*, pages 223–255.

Ina Rösiger. 2018. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Katrin Schweitzer, Kerstin Eckart, Markus Gärtner, Agnieszka Falenska, Arndt Riester, Ina Rösiger, Antje Schweitzer, Sabrina Stehwien, and Jonas Kuhn. 2018. German radio interviews: The GRAIN release of the SFB732 silver standard collection. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Joseba Rodríguez, and Massimo Poesio. 2019. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26:95 – 128.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Amir Zeldes. 2022. Opinion piece: Can we fix the scope for coreference? *Dialogue & Discourse*, 13:41–62.

A GUMBridge Bridging Subtypes

This appendix briefly details the bridging subtype varieties annotated in the GUMBridge corpus, which are reflected in its guidelines and in the prompts used in the GUMBridge paper to elicit LLM annotations. For the full guidelines, we refer readers to Levine and Zeldes 2026a.

COMPARISON-RELATIVE The anaphor is preceded by a comparative marker which implies a comparison to the antecedent (e.g., several women → **other women**).

COMPARISON-SENSE The type of the anaphor is omitted but inferable via comparison to the antecedent (e.g., a Chinese restaurant → **the Italian one**).

COMPARISON-TIME The anaphor refers to a specific time/time frame which is understandable with reference to the time/time frame expressed by the antecedent (e.g., Wednesday → **yesterday**).

ENTITY-MERONOMY The anaphor has a part-whole relation with the antecedent, including physical subparts, substance-portion, and regions/subsections (e.g., a house → **the door**).

ENTITY-PROPERTY The anaphor is a physical or intangible property of the antecedent, such as smell, length, or style (e.g., a bouquet of roses → **the scent**).

ENTITY-RESULTATIVE The anaphor is logically inferable from the antecedent. This is often the result of a transformative/product producing process, like cooking/baking (e.g., flour → **the bread**).

ENTITY-ASSOCIATIVE The anaphor is an attribute or closely associated entity of the antecedent (e.g., a library → **the books**).

SET-MEMBER The anaphor is an element of the antecedent set. This includes group-member and class-instance relations (e.g., several books → **the mystery novel**).

SET-SPAN-INTERVAL The anaphor is a sub-span of the spatial or temporal antecedent interval (e.g, Sunday → **the morning**).

SET-SUBSET The anaphor is a subset of the antecedent set (e.g., a group of students → **the boys**).

OTHER The OTHER category is for instances which fit the information status based definition of a bridging pair but do not fall into any of the bridging subtype categories outlined above.

B Bridging Subtype Label Counts in Error Classes

Table 3 gives the raw counts of the bridging subtype labels that occur in each error class (including True Positives) for the human and LLM annotation of bridging instances. No False Positive counts are provided for the LLM annotations because the LLM pipeline was divided into subtasks (anaphor recognition, antecedent resolution, and subtype classification), and the LLMs were not queried on subtype classification for False Positives.

Subtype	Human			LLM	
	TP	FP	FN	TP	FN
COMPARISON-RELATIVE	53	32	70	35	104
COMPARISON-SENSE	26	8	34	19	45
COMPARISON-TIME	17	36	28	1	49
ENTITY-ASSOCIATIVE	118	104	305	25	198
ENTITY-MERONOMY	29	24	24	32	30
ENTITY-PROPERTY	4	9	19	0	30
ENTITY-RESULTATIVE	8	17	11	1	27
SET-MEMBER	16	51	19	9	42
SET-SPAN-INTERVAL	7	20	7	2	18
SET-SUBSET	11	30	12	1	18
OTHER	1	11	4	1	8

Table 3: Human and LLM subtype label counts in the different error classes: True Positives (TP), False Positives (FP), and False Negatives (FN).

C LLM Pipeline and Prompts

In this appendix we provide a brief description of the LLM pipeline used to create the bridging resolution baseline data analyzed in this paper. Please see Levine and Zeldes 2026a for further details.

The LLMs are queried individually for each of the following bridging resolution subtasks: (1) anaphora recognition, (2) antecedent selection, and (3) subtype categorization. For each subtask, the models are provided with a separate prompt which gives a detailed explanation of the task based on the GUMBridge annotation guidelines, with a series of few-shot examples. For the anaphor recognition subtask, the models are queried sentence by sentence through the document. For the antecedent selection subtask, the models are queried once for each bridging anaphor in the gold annotations. For the subtype categorization subtask, the models are queried once for each bridging pair in the gold annotations. Prompt templates for each subtask are included below.

C.1 Anaphor Recognition

You are a linguistic analyst whose job is to find cases of bridging anaphora: mentions of newly introduced entities (noun phrases) in a text, for which a reader would need to refer back to a previously mentioned, non-identical entity to resolve their meaning. There are several classes of bridging anaphors, any of which should be identified in the text being analyzed. In the following examples, the bridging anaphor is surrounded by *asterisks*.

comparison-relative: The anaphor is preceded by a comparative marker (other, another, same, more, ordinal modifiers, comparative adjectives, superlatives, etc.) which implies a comparison to the antecedent. For example: "The children... *another child*" (=another with comparison to the aforementioned children); similar cases may be *similar children*, *older children* (compared to the aforementioned children), etc.

comparison-sense: the semantic type of a phrase requires a previous mention to identify it, for example "the Italian "estaurant... *a Chinese one*" (we can't know "a Chinese one" is a restaurant without referring back to the Italian restaurant), or "*another one*", "*the others*" etc.

comparison-time: the anaphor refers to a specific time/timeframe which is understandable with reference to the antecedent, for example: "Tuesday, February 2nd ... *the following week*"

entity-meronomy: the anaphor is a subunit of the antecedent (part-whole), including physical subunits, portion-substance relations, and regions/subsections. For example: "the house ... *the door*" (=of the house).

entity-associative: the anaphor is an attribute or closely associated entity of the antecedent, including both prototypical and inducible associations: "a wedding ... *the bride*" (=the bride at that wedding), implicit arguments of a predicate or a verbal nominalization: "a play... *the performance*" (=of the play), relational nouns: "a murder ... *the victim*"

entity-property: the anaphor is a physical or intangible property of the antecedent (e.g., smell, length, size, style, etc.): "the tea... *the sweet aroma*"

entity-resultative: the anaphor is logically inferable from the antecedent (e.g., result, transformation/transmutation, cause): "the dough ... *the bread*" (=the dough becomes bread after baking)

set-member: the anaphor is an element of the antecedent set, including groups-member relations and classes-instances: "the cars ... *the Mazda*", additionally indefinite members to definite sets: "a candle on each cupcake... *the candles*"

set-subset: the anaphor is a subset of the antecedent set: "the cars ... *the Mazdas*" (not all Mazdas, just the subset among the aforementioned cars)

set-span-interval: the anaphor is a sub-span of a spatial or temporal interval defined by the antecedent: "last week... *Wednesday*" (=Wednesday of last week), "Sunday... *the morning* (=the morning portion of that Sunday)"

other: the anaphor requires a previous entity for interpretation, but it doesn't fit into any of the above categories. This is a rare class.

There are also some exceptions which should NOT be identified as bridging anaphora:

Coreference: If an entity has a previous mention, it cannot be an instance of bridging. For instance, in "Catherine and Henry had their wedding last week. The bride was very beautiful", even though there is an associative relationship between the wedding and the bride, since "the bride" corefers with "Catherine", which has already been introduced to the discourse, "the bride" is not eligible to be an instance of bridging.

Bridging-contained: If the entity one would need to refer back to in order to understand the bridging anaphor is a direct modifier in the noun phrase of the potential bridging anaphor, e.g. "the focus of the story" or "two of them", it should not be annotated as bridging. In other words, the previous antecedent entity must be outside of the nominal phrase containing the anaphor. An entity that is followed by a prepositional phrase or a relative clause is sufficiently qualified and is thus NOT an instance of bridging.

Generics/Situational bridging: Entities that are accessible due to general world knowledge or situational context are not considered instances of bridging, i.e., if it doesn't have a previous associated antecedent entity to be bridging from, it cannot be bridging.

Possession with an explicit possessive: If the potential bridging anaphor contains an explicit possessive which corefers with the associative antecedent entity, no bridging relation is necessary. Explicit coreference between the associative antecedent and the possessive is sufficient (e.g., [Mark]. . . [his] house → no bridging, coreference between "Mark" and "his"). Contrast this with [the family] . . . *the house* → bridging, since we cannot interpret which house it is (the house of the family) without referring to "the family", which is outside of the anaphor phrase.

Here are 2 an examples of the task:

Please return a list all of the bridging anaphors in the following text in the order in which they appear. Output the anaphor mention phrase exactly as it appears in the text. If there are no bridging anaphors, return an empty list.

Text:

... with their friends to a picnic. The picnic was supposed to take place in a grove, but the shade wasn't enough, so they had to find a different place. Conny started to say ...

Answer(s):

["the shade", "a different place"]

Please return a list all of the bridging anaphors in the following text in the order in which they appear. Output the anaphor mention phrase exactly as it appears in the text. If there are no bridging anaphors, return an empty list.

Text:

... making this technique the basis of training for all types of dance . While dancing ballet takes dedication and requires serious training , you can learn the basics to prepare yourself for further study . Learn to get ready for practicing...

Answer(s):

["the basics", "further study"]

Please return a list all of the bridging anaphors in the following text in the order in which they appear. Output the anaphor mention phrase exactly as it appears in the text. If there are no bridging anaphors, return an empty list.

Text:

{text}

Answer(s):

C.2 Antecedent Selection

You are a linguistic analyst whose job is to select the associative antecedent for of a bridging anaphor: mentions of newly introduced entities (noun phrases) in a text, for which a reader would need to refer back to a previously mentioned, non-identical entity (the antecedent) to resolve their meaning. There are several classes of bridging instances, defined by the associative relationship between the bridging anaphor and its associative antecedent. In the following examples, the bridging antecedent is surrounded by *asterisks* and the bridging anaphor is surrounded by {{double curly brackets}}.

comparison-relative: The anaphor is preceded by a comparative marker (other, another, same, more, ordinal modifiers, comparative adjectives, superlatives, etc.) which implies a comparison to the antecedent. For example: "*The children* ... {{another child}}" (=another with comparison to the aforementioned children); similar cases may be {{similar children}}, {{older children}}(compared to the aforementioned children), etc.

comparison-sense: the semantic type of a phrase requires a previous mention to identify it, for example "*the Italian restaurant* ... {{a Chinese one}}" (we can't know "a Chinese one" is a restaurant without referring back to the Italian restaurant), or "{{another one}}", "{{the others}}" etc.

comparison-time: the anaphor refers to a specific time/timeframe which is understandable with reference to the antecedent, for example: "*Tuesday, February 2nd* ... {{the following week}}"

entity-meronomy: the anaphor is a subunit of the antecedent (part-whole), including physical subunits, portion-substance relations, and regions/subsections. For example: "*the house* ... {{the door}}" (=of the house).

entity-associative: the anaphor is an attribute or closely associated entity of the antecedent, including both prototypical and inducible associations: "*a wedding* ... {{the bride}}" (=the bride at that wedding), implicit arguments of a predicate or a verbal nominalization: "*a play*... {{the performance}}" (=of the play), relational nouns: "*a murder* ... {{the victim}}"

entity-property: the anaphor is a physical or intangible property of the antecedent (e.g., smell, length, size, style, etc.): "*the tea* ... {{the sweet aroma}}"

entity-resultative: the anaphor is logically inferable from the antecedent (e.g., result, transformation/transmutation, cause): "*the dough* ... {{the bread}}" (=the dough becomes bread after baking)

set-member: the anaphor is an element of the antecedent set, including groups-member relations and classes-instances: "*the cars* ... {{the Mazda}}", additionally indefinite members to definite sets: "*a candle*

on each cupcake ... {{the candles}}"

set-subset: the anaphor is a subset of the antecedent set: "*the cars* ... {{the Mazdas}}" (not all Mazdas, just the subset among the aforementioned cars)

set-span-interval: the anaphor is a sub-span of a spatial or temporal interval defined by the antecedent: "*last week* ... {{Wednesday}}" (=Wednesday of last week), "*Sunday* ... {{the morning}}" (=the morning portion of that Sunday)

other: the anaphor requires a previous entity for interpretation, but it doesn't fit into any of the above categories. This is a rare class.

Here are 2 an examples of the task:

Please return a single string for associative antecedent of the bridging anaphor surrounded by {{double curly brackets}}. Output the antecedent mention phrase exactly as it appears in the text. If there is no associative antecedent, return "no antecedent". The antecedent you are returning CANNOT be the same as the bracketed anaphor.

Text:

... with their friends to a picnic. The picnic was supposed to take place in a grove, but {{the shade}} wasn't enough, so they had to find a different place. Conny started to say ...

Answer:

a grove

Please return a single string for associative antecedent of the bridging anaphor surrounded by {{double curly brackets}}. Output the antecedent mention phrase exactly as it appears in the text. If there is no associative antecedent, return "no antecedent".

Text:

... making this technique the basis of training for all types of dance . While dancing ballet takes dedication and requires serious training , you can learn the basics to prepare yourself for {{further study}} . Learn to get ready for practicing ...

Answer:

ballet

Please return a single string for associative antecedent of the bridging anaphor surrounded by {{double curly brackets}}. Output the antecedent mention phrase exactly as it appears in the text. If there is no associative antecedent, return "no antecedent".

Text:

{text}

Answer:

C.3 Subtype Categorization

You are a linguistic analyst whose job is to select the subtype classification for a bridging anaphor - antecedent pair: mentions of newly introduced entities (the anaphor) in a text, for which a reader would need to refer back to a previously mentioned, non-identical entity (the antecedent) to resolve their meaning. There are several classes of bridging instances, defined by the associative relationship between the bridging anaphor and its associative antecedent. In the following subtype examples, the bridging antecedent is surrounded by *asterisks* and the bridging anaphor is surrounded by {{double curly brackets}}.

comparison-relative: The anaphor is preceded by a comparative marker (other, another, same, more, ordinal modifiers, comparative adjectives, superlatives, etc.) which implies a comparison to the antecedent. For example: "*The children* ... {{another child}}" (=another with comparison to the aforementioned children); similar cases may be {{similar children}}, {{older children}} (compared to the aforementioned children), etc.

comparison-sense: the semantic type of a phrase requires a previous mention to identify it, for example "*the Italian restaurant* ... {{a Chinese one}}" (we can't know "a Chinese one" is a restaurant without referring back to the Italian restaurant), or "{{another one}}", "{{the others}}" etc.

comparison-time: the anaphor refers to a specific time/timeframe which is understandable with reference to the antecedent, for example: "*Tuesday, February 2nd* ... {{the following week}}"

entity-meronomy: the anaphor is a subunit of the antecedent (part-whole), including physical subunits, portion-substance relations, and regions/subsections. For example: "*the house* ... {{the door}}" (=of the house).

entity-associative: the anaphor is an attribute or closely associated entity of the antecedent, including both prototypical and inducible associations: "*a wedding* ... {{the bride}}" (=the bride at that wedding), implicit arguments of a predicate or a verbal nominalization: "*a play*... {{the performance}}" (=of the play), relational nouns: "*a murder* ... {{the victim}}"

entity-property: the anaphor is a physical or intangible property of the antecedent (e.g., smell, length, size, style, etc.): "*the tea* ... {{the sweet aroma}}"

entity-resultative: the anaphor is logically inferable from the antecedent (e.g., result, transformation/transmutation, cause): "*the dough* ... {{the bread}}" (=the dough becomes bread after baking)

set-member: the anaphor is an element of the antecedent set, including groups-member relations and classes-instances: "*the cars* ... {{the Mazda}}", additionally indefinite members to definite sets: "*a candle* on each cupcake ... {{the candles}}"

set-subset: the anaphor is a subset of the antecedent set: "*the cars* ... {{the Mazdas}}" (not all Mazdas, just the subset among the aforementioned cars)

set-span-interval: the anaphor is a sub-span of a spatial or temporal interval defined by the antecedent: "*last week* ... {{Wednesday}}" (=Wednesday of last week), "*Sunday* ... {{the morning}}" (=the morning portion of that Sunday)

other: the anaphor requires a previous entity for interpretation, but it doesn't fit into any of the above categories. This is a rare class.

Here are 2 an examples of the task:

In the following text, a bridging anaphora is marked with {{double curly brackets}} and the corresponding antecedent is surrounded by *asterisks*. Read the following text and for the bridging anaphor-antecedent pair, classify the variety of bridging subtype relation (defined above) that holds between the two entities. Multiple subtypes may apply to a single pair. Output a string of all applicable subtypes, connected by semicolons (no spaces).

The possible subtype labels are as follows:
comparison-relative
comparison-sense
comparison-time
entity-associative
entity-meronomy
entity-property
entity-resultative
set-member
set-subset
set-span-interval
other

Antecedent Text:

... with their friends to a picnic. The picnic was supposed to take place in *a grove*, but the shade wasn't enough, so they had to find a different place. Conny started to say ...

Anaphor Text:

... to a picnic. The picnic was supposed to take place in a grove, but {{the shade}} wasn't enough, so they had to find a different place. Conny started to say ...

Answer:

entity-associative

In the following text, a bridging anaphora is marked with {{double curly brackets}} and the corresponding antecedent is surrounded by *asterisks*. Read the following text and for the bridging anaphor-antecedent pair, classify the variety of bridging subtype relation (defined above) that holds between the two entities. Multiple subtypes may apply to a single pair. Output a string of all applicable subtypes, connected by semicolons (no

spaces).

The possible subtype labels are as follows:
comparison-relative
comparison-sense
comparison-time
entity-associative
entity-meronomy
entity-property
entity-resultative
set-member
set-subset
set-span-interval
other

Antecedent Text:

... this technique the basis of training for all types of dance . While dancing *ballet* takes dedication and requires serious training , you can learn the basics to prepare yourself for further study . Learn to get ready for practicing ...

Anaphor Text:

... making this technique the basis of training for all types of dance . While dancing ballet takes dedication and requires serious training , you can learn the basics to prepare yourself for {{further study}} . Learn to get ready for ...

Answer:

comparison-relative

In the following text, a bridging anaphora is marked with {{double curly brackets}} and the corresponding antecedent is surrounded by *asterisks*. Read the following text and for the bridging anaphor-antecedent pair, classify the variety of bridging subtype relation (defined above) that holds between the two entities. Multiple subtypes may apply to a single pair. Output a string of all applicable subtypes, connected by semicolons (no spaces).

The possible subtype labels are as follows:
comparison-relative
comparison-sense
comparison-time
entity-associative
entity-meronomy
entity-property
entity-resultative
set-member
set-subset
set-span-interval
other

Antecedent Text:

... {antecedent_text} ...

Anaphor Text:

... {anaphor_text} ...

Answer: