Not All Disneys Are the Same: Making Coreference Metonymy-Aware

Bingyang Ye, Jingxuan Tu, James Pustejovsky
Michtom School of Computer Science, Brandeis University
415 South Street, Waltham, MA 02453, USA
{byye, jxtu, jamesp}@brandeis.edu

Abstract

Metonymy, a type of referential transfer in which a name evokes a conceptually related entity (e.g., "Disney" for the theme park), is a pervasive and systematic feature of natural language. Yet, despite its impact on entity interpretation, coreference research has rarely treated metonymy explicitly. Computational models of metonymy, in turn, typically analyze local, sentence-level cases, leaving unexplored how metonymic reference interacts with discourse-level coreference phenomena. We bridge this gap by introducing CoNLL-Coref-Met, a metonymy-aware annotation layer on top of CoNLL-2012 that flags metonymic mentions in context. Using this lens, we show that state-of-the-art neural resolvers and LLMs systematically underperform on metonymic clusters relative to literal counterparts. We then (i) correct clusters affected by metonymy to reflect semantic reference rather than surface form and (ii) introduce a metonymy-aware LLM procedure to resolve semantic ambiguities introduced by metonymic shifts. Our pipeline introduces a novel way to see, measure, and mitigate metonymy effects on coreference.

Keywords: coreference, semantics, corpus

1. Introduction

Coreference resolution (CR) is typically framed as an identity task: two mentions corefer if they denote the same discourse entity. This framing has supported steady progress and fair comparison across systems (Pradhan et al., 2012; Bamman et al., 2020; Webster et al., 2018; Ghaddar and Langlais, 2016). However, it is challenged by figurative language, especially metonymy, where a single name evokes different but closely related facets of an entity in context (Pustejovsky and Rumshisky, 2009; Ye et al., 2022; Johnson and Lakoff, 1980; Fauconnier, 1994; Ježek, 2016). In newswire and conversational text, writers routinely use names such as *France* for its government, *Harvard* for the institution or its people, or *Wall Street* for the financial sector. These sense shifts complicate coreference because surface-identical names do not always align straightforwardly with discourse identity.

Consider the following examples:

(1) **[Washington]** believes North Korea may have enough nuclear fuel to make more than eight or nine atomic weapons… **[The U.S.]** continues to push for disarmament in the region.

(2) **[The Bush administration]** maintains that the President has **[the country's]** best interests at heart. This is Kelli Arena in **[Washington]**, reporting live from **[the White House]**.

In (1), *Washington* refers to the U.S. government rather than the geographic location, and therefore belongs in the same discourse chain as *The U.S.*. By contrast, in (2), *Washington* is used literally as a location and should not corefer with *the country*, while *the White House* should not corefer with *The Bush administration*. These examples illustrate a central challenge for CR: systems must determine not only whether mentions are lexically or semantically related, but also whether they refer to the same discourse entity under the intended contextual reading.

Although the OntoNotes guidelines[1] acknowledge several metonymic patterns and provide instructions for annotators, the effect of metonymy on coreference evaluation has been largely overlooked. When such facet shifts occur, identity-only modeling can become brittle: systems may overmerge related but distinct entities, or fragment a single discourse referent across its surface facets. As a result, standard coreference evaluation can obscure whether models are clustering mentions based on intended meaning or merely on surface similarity. This blind spot is especially relevant for OntoNotes-style newswire, where metonymic uses of organization and location names are common.

A complementary line of work has focused on metonymy resolution itself, proposing taxonomies, detectors, and benchmarks for distinguishing literal from metonymic readings, especially for organization and location mentions (Markert and Nissim, 2007). However, these resources are typically sentence-level, rather than document-level, and therefore provide limited evidence about how metonymy interacts with downstream discourse

[1] https://ufal.mff.cuni.cz/pcedt3.0/pubs/english-coreference-guidelines.pdf

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 12019–12030
11–16 May 2026.

tasks such as coreference (Ghosh and Jiang, 2025). We still lack a benchmark that makes it possible to study metonymy inside document-level coreference, where discourse context determines whether a facet shift should be integrated into an entity chain or kept separate.

To address this gap, we introduce a human-verified, metonymy-aware annotation layer over the English CoNLL-2012 development set (Pradhan et al., 2012), focusing on organization and location mentions, where facet shifts are especially frequent and consequential. The layer is intentionally lightweight: it preserves the original mention spans and coreference chains while adding labels that identify where metonymy changes mention interpretation. This allows us to evaluate existing coreference systems under metonymy-sensitive conditions while retaining comparability with a widely used benchmark.

Using this layer, we examine three questions: (1) how much current coreference systems degrade when metonymic mentions are present; (2) whether different models exhibit different error profiles on mentions that should be merged versus separated under metonymic interpretation; and (3) whether explicit metonymy cues can improve document-level coreference clustering. Our analysis shows that both neural resolvers and LLM-based systems underperform on metonymy-bearing clusters relative to literal ones. Their errors follow two recurring patterns: over-merges, where related but distinct entities are collapsed into one chain, and fragmentation, where one discourse entity is split across multiple surface facets.

Beyond diagnosis, we also explore a simple mitigation strategy. We propose a metonymy-aware, contextual cue-guided LLM pipeline in which the model first determines whether a mention is used literally or metonymically, and then performs document-level clustering using those judgments as contextual cues. This setup leaves final linking decisions to the document context while helping the model avoid some of the over-merges and fragmentation triggered by unresolved facet shifts.

In summary, this paper makes three contributions. First, we introduce a metonymy-aware annotation layer for the English CoNLL-2012 development set. Second, we use it to evaluate modern coreference systems and show that metonymic mentions remain a systematic source of errors. Third, we present a simple metonymy-aware prompting strategy that yields consistent numerical improvements in document-level clustering. By bringing metonymy into the coreference evaluation loop, we provide a concrete way to study figurative language within a standard benchmark setting.[2]

[2]https://github.com/brandeis-llc/met_anaphor.

2. Related Works

2.1. Coreference Resolution

Coreference Resolution has progressed from feature-engineered pipelines to neural, end-to-end systems that jointly model mention detection and clustering (Lee et al., 2017). Span-based architectures introduced the now-standard paradigm of enumerating spans, scoring them as mentions, and linking via antecedent selection with global regularization; these models achieved strong results on OntoNotes/CoNLL-2012 and helped standardize evaluation with MUC, B^3, and CEAF, which is often reported as the CoNLL F1 (Pradhan et al., 2012). Subsequent work improved mention proposal with better span pruning, learned width priors, and coarse-to-fine pruning for efficiency on long documents (Lee et al., 2018). Contextual encoders substantially lifted performance by providing richer token representations and by enabling cross-sentence reasoning; further gains came from entity-aware encoders that inject learned entity markers and global features such as cluster size, entity type, and discourse salience (Wu et al., 2020).

Beyond single-document CR, cross-document and event/entity linking frameworks extend clustering across collections—e.g., cross-document entity coref with distributed inference and hierarchical models (Singh et al., 2011; Cattan et al., 2021; Allaway et al., 2021) and cross-document event coref/search (Eirew et al., 2022; Chen et al., 2023); these are often coupled with entity linking to KBs (Finin et al., 2009; Shen et al., 2014). Specialized settings—pronoun resolution (Webster et al., 2018; Chada, 2019), nested/overlapping mentions handled via span enumeration (Lee et al., 2017), and document-level long-range dependencies with memory/hierarchical encoders or long-context Transformers (Kantor and Globerson, 2019; Xia et al., 2020; Beltagy et al., 2020)—spurred methods that combine span models with global memory or hierarchical structure.

Recently, large language models (LLMs) have reshaped the design space. Instruction-following LLMs can perform "clustering-only" CR when given gold spans, or end-to-end CR via structured prompting that enforces span indices and cluster IDs (Le and Ritter, 2023). QA-style formulations offer controllable precision/recall trade-offs (Wu et al., 2020); chain-of-thought or rationale-augmented prompts can expose model uncertainty (Wei et al., 2022). Hybrid systems use LLMs for difficult decisions while retaining neural span scorers for coverage and speed. Practical advances include domain adaptation (Uzuner et al., 2012), multilingual CR (Skachkova, 2024), and robustness studies that probe sensitivity to prompt wording (Zhao et al., 2021; Lu et al., 2022), and distribution shift (Le and

Ritter, 2023)—all increasingly relevant when CR is embedded in larger IE/QA/agentic pipelines.

2.2. Metonymic Resolution

Metonymy uses a referential expression to stand for a related entity. It sits at the intersection of lexical semantics and reference. Classic linguistic accounts of metonymy, such as Generative Lexicon, often involve type coercion over qualia structure and logical polysemy (Pustejovsky, 1995; Pustejovsky and Rumshisky, 2009; Romani and Ježek, 2020). Conventionalized patterns like PLACE→PEOPLE or ORG→MEMBERS/LOCATION/PRODUCT were operationalized in shared-task guidelines for SemEval-2007 Task 08 (Markert and Nissim, 2007). Early computational approaches based on rules, selectional preferences, and hand-crafted features evolved into supervised classifiers that use distributional and syntactic cues (Markert and Nissim, 2002; Nissim and Markert, 2003).

Neural methods adapt these ideas with contextual encoders, casting metonymy recognition as (i) binary or metatype classification for a marked span or (ii) multi-label mapping to fine-grained "target" types. Word or span level BERT models for location metonymy exemplify this trend (Li et al., 2020), and adding structural constraints further improves robustness (Wang et al., 2023). These models leverage syntactic roles, predicate semantics, and selectional constraints; sequence-labeling variants treat metonymy like a specialized NER layer, while span-typing aligns naturally with referential shifts.

More recent research in metonymy leverages LLM to recoginize metonymic patterns. A recent dataset and method along these lines is ConMeC for common-noun metonymy (Ghosh and Jiang, 2025). Yet LLMs remain brittle under domain shift and figurative density, motivating diagnostics that partition clusters by metonymic content and measure degradation relative to literal-only clusters.

3. Data Annotation

We create CoNLL-Coref-Met, a dataset that augments the CoNLL-2012 coreference corpus with metonymy labels through an LLM-assisted screening step followed by human annotation and adjudication. The underlying CoNLL-2012 dataset is the prevailing benchmark for document-level coreference. Due to limited resources, we focus on the English development set to create a controlled evaluation benchmark under limited annotation resources; extending the layer to train/test splits is an important next step.

Following Markert and Nissim (2007), we annotate two semantic classes of mentions, Locations and Organizations, where facet shifts due to metonymy are frequent and consequential. Each target mention is annotated for both metonymy status (LITERAL vs. METONYMIC) and cluster compatibility, yielding four cluster-aware labels. This taxonomy offers well-defined, historically validated categories for metonymy in named entities.

3.1. Candidate Cluster Identification

Because our taxonomy only focuses on locations and organizations, we first filter the dataset to select candidate clusters. Specifically, we run an NER pipeline over all the documents using spaCy and we retain only those clusters that contain at least one mention whose entity type is an organization or location. Clusters are kept intact even if they include pronominal or nominal references. For example, if a cluster contains [*HongKong Disney, Disneyland, it*], we keep the entire cluster, including the pronoun *it*, because the cluster is referring to a location.

3.2. LLM-Assisted Metonymy Resolution

For every mention in the candidate clusters, we run an LLM screening step to flag potential metonymic uses. Our prompting strategy takes inspiration from ConMeC (Ghosh and Jiang, 2025), which uses a two-step, category-first prompting scheme and chain-of-thought rationales: (i) assign the semantic category of the target mention; and (ii) apply a category-dependent checklist to decide metonymic vs. literal. In our case, we instantiate only the ORGANIZATION and LOCATION branches. We replicate this logic but constrain it to our two target categories, using `GPT-5` for the screening step. This LLM step is used only for screening and pre-flagging potential metonymic mentions; it does not determine the final dataset labels. Because LLM suggestions may introduce anchoring effects in borderline cases, all final labels are assigned by human annotators, and we report inter-annotator agreement and adjudication results below.

3.3. Mention Annotation

We then manually validate all target mentions in the candidate clusters and assign the final cluster-aware labels. Each mention receives one of four labels that disentangle referential shift from clustering quality:

1. LITERAL-IN-CLUSTER — literal mention correctly placed in its gold cluster
2. LITERAL-OUT-OF-CLUSTER — literal mention that should belong to a different cluster
3. METONYMIC-IN-CLUSTER — metonymic mention correctly clustered with its intended referent

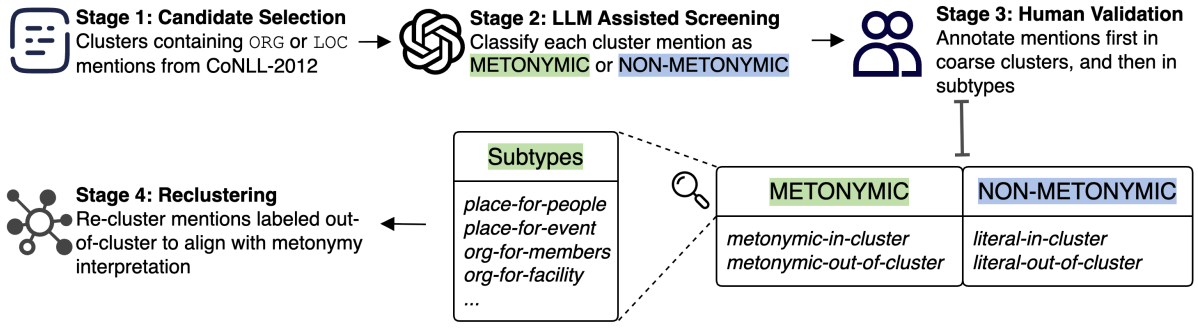

Figure 1: Our annotation workflow.

Documents	94
Clusters	204
Mentions	859
Docs with metonymy	50 (53.2%)
Clusters with metonymy	73 (35.8%)
Metonymy rate (mentions)	20.1%
Cluster categories	
LITERAL	131 (64.2%)
MET_IN	22 (10.8%)
MET_OUT	51 (25.0%)

Table 1: Dataset statistics for CoNLL-Coref-Met.

4. METONYMIC-OUT-OF-CLUSTER — metonymic mention that should be clustered elsewhere

We also manually annotate the metonymy subtype (e.g., place-for-people, org-for-members, etc.) in Markert and Nissim (2007) for future model error analysis. We build a web application as our annotation tool that presents the mention in the discourse context and enforces per-mention labeling while keeping the cluster visible, which speeds annotation and reduces context switching. We also evaluate the LLM screening step against the final human annotations; it achieves an F1 of 76.91%.

3.4. Statistics

The annotated development set contains 94 documents with 204 coreference clusters and 859 total mentions. Over half of the documents contain at least one metonymic mention, and roughly one third of all clusters are affected by metonymy. Overall, 20.1% of mentions exhibit metonymic usage. Among clusters, 64.2% are purely literal, while 10.8% include metonymic mentions within the same cluster, and 25.0% correspond to separate metonymic readings. These figures indicate that metonymy is both frequent and structurally diverse in location and organization mentions, making it a meaningful source of variation for evaluating coreference models. We report the statistics of our dataset in Table 1.

3.5. Reclustering

Finally, we create CoNLL-Coref-Met-Reclustered where annotators manually reclustered all mentions in CoNLL-Coref-Met which are marked as out-of-cluster, whether literal or metonymic, into their appropriate referential groups, producing a corrected set of clusters aligned with metonymic interpretations. The annotators are asked to assign the out-of-cluster mention with the correct cluster id. If there is no existing cluster the mention belongs to then a new cluster should be generated and the mention should be linked to the new cluster. This reclustering fully resolved all 119 out-of-cluster mentions. The resulting merged annotations yield a more complete and internally consistent representation of reference, forming a parallel *metonymy-aware* gold standard for evaluating coreference models under metonymic shift.

3.6. Annotator Agreement

Two graduate students in Computer Science at a U.S.-based university conducted the annotations. All candidate mentions, including mention subtypes, and reclustering decisions were independently annotated by the two annotators. Disagreements were adjudicated through discussion. Inter-annotator agreement is 85.6% (κ) for the cluster-aware metonymy labels, and 76.3% on the reclustering decisions. Inter-annotator agreement is 85.6% (κ) for the cluster-aware metonymy labels, and 76.3% agreement on the reclustering decisions. The lower agreement for reclustering reflects the fact that this step requires global discourse-level judgments about entity reference, rather than local classification of metonymy at the mention level. Most disagreements occurred in borderline cases where multiple related entities (e.g., a country vs. its government) could plausibly be interpreted as sharing or separating a discourse referent depending on context. All such disagreements were resolved through adjudication, and the released dataset

uses the adjudicated clusters as the final gold standard.

The pipeline substantially reduces human effort while producing metonymy-aware coreference data. It also preserves full discourse chains, enabling analyses of how metonymy affect clustering errors and evaluation metrics in a controlled, comparable setting. We show the full annotation pipeline in Figure 1.

4. Are Current Coreference Resolvers Sensitive to Metonymy?

We investigate whether contemporary coreference resolvers are systematically sensitive to metonymy. Our starting point is CoNLL-Coref-Met, which we annotate with mention-level metonymy annotations. Each gold mention is labeled as literal or metonymic. These labels allow us to partition gold clusters into analytically distinct environments. By comparing model behavior across these environments, we can evaluate model performance shift driven by metonymic shifts within a cluster.

4.1. Evaluation design

To examine model sensitivity to metonymy, we evaluate coreference performance on three subsets of gold clusters rather than on the entire document. Each gold cluster is first categorized by the metonymy status of its mentions: (i) Literal clusters contain only literal mentions, (ii) Metonymic-in-cluster (Met_In) clusters include one or more metonymic mentions that legitimately belong to the same entity (e.g., *the White House* referring to the administration), and (iii) Metonymic-out-of-cluster (Met_Out)clusters correspond to cases where a metonymic mention refers to a distinct entity and thus should not corefer with the literal one. We then compute scores separately for each subset to quantify how model behavior changes when metonymy is involved. Evaluating by cluster category allows us to isolate metonymy-induced errors—splits for Met_In and spurious merges for Met_Out—and to test whether current resolvers distinguish literal from metonymic readings. These subset-level analyses serve as a sensitivity test, highlighting how figurative reference affects coreference grouping.

4.2. Evaluation Metrics

We report metonymy sensitivity using cluster-based B^3 and LEA macro-averaged over gold clusters.

Let each gold cluster be denoted as G and each predicted cluster as P. For every mention $m \in G$, let $P(m)$ be the predicted cluster containing m and $G(m)$ be its gold cluster.

Cluster-based B^3. Following Bagga (1998), we compute B^3 precision and recall for each gold cluster while keeping the full predicted cluster size to avoid masking out-of-cluster false positives:

$$\text{Prec}(G) = \frac{1}{|G|} \sum_{m \in G} \frac{|P(m) \cap G|}{|P(m)|}, \quad (1)$$

$$\text{Rec}(G) = \frac{1}{|G|} \sum_{m \in G} \frac{|P(m) \cap G|}{|G|}. \quad (2)$$

The per-cluster F1 is

$$\text{F}_1(G) = \frac{2 \times \text{Prec}(G) \times \text{Rec}(G)}{\text{Prec}(G) + \text{Rec}(G)}, \quad (3)$$

and macro-averaged across all gold clusters of a given category, i.e., Literal, Metonymic-IN, Metonymic-OUT.

Cluster-based LEA. LEA measures how well intra-cluster links in the gold data are preserved in prediction (Moosavi and Strube, 2016). Let $L(G)$ be the set of unordered mention pairs (links) in G:

$$L(G) = \{ (m_i, m_j) \mid m_i, m_j \in G, i < j \}. \quad (4)$$

For each gold cluster G, its LEA resolution score is defined as

$$\text{LEA}(G) = \frac{|L(G) \cap L(P)|}{|L(G)|}, \quad (5)$$

where $L(P)$ is the set of predicted links. The overall cluster-based LEA is the macro-average of LEA(G) across clusters in each category. This formulation directly penalizes splits of a gold entity missing intra-cluster links and merges with incorrect mentions, providing a link-level view of model sensitivity to metonymy.

We adopt these metrics for two reasons. First, cluster-restricted variants of CoNLL-2012 (MUC/B^3/CEAF) are not well-defined: scoring on a subset of clusters drops predicted links that cross the subset boundary, which *masks* precision errors and inflates scores, especially when metonymic clusters are rare. In contrast, our cluster-level B^3 keeps each predicted cluster's full size $|P(m)|$ when computing precision, so spurious merges to out-of-cluster mentions still penalize the score, and LEA directly measures preservation of intra-cluster links, making splits (common for metonymic readings) visible. Second, macro-averaging at the cluster level reduces document-level confounds and allows fair comparison across categories (LITERAL, MET_IN, MET_OUT), while size-bucket analyses further control for cluster-size effects. For these reasons, cluster-based B^3 and LEA provide a faithful, diagnostics-oriented view of how metonymy impacts coreference, without the masking artifacts introduced by subset CoNLL scoring.

For completeness and comparability with prior work, we also report document-level CoNLL F1 on the same three subsets of documents, i.e., those containing only literal clusters, containing any METONYMIC-IN clusters, and containing any METONYMIC-OUT clusters. Unlike our cluster-based analysis, this evaluation preserves the full document context and does not mask links to out-of-subset mentions, thereby reflecting the model's overall end-to-end performance under varying degrees of metonymy presence. Reporting document-level CoNLL F1 alongside cluster-based metrics enables direct comparison to standard benchmarks while revealing how metonymic content affects global entity resolution within realistic document boundaries.

4.3. Experiments

We evaluate our dataset with SOTA CR system `Maverick` (Martinelli et al., 2024) and LLM-based systems. To control for mention-detection confounds, all systems operate over the same oracle mentions. For `Maverick`, we use the `maverick-mes-ontonotes` checkpoint which is trained on the training set of OntoNotes data with clustering-only mode where the model are given the predefined gold mentions. For LLM systems, we use `Qwen 2.5-32B-Instruct` and `GPT-4.1` model. And we follow the formulation introduced in Le and Ritter (2023), where the language model is prompted with a *document template* that marks each candidate mention in the text and instructs the model to annotate them with cluster identifiers directly within the passage. This document-level prompting allows the model to jointly reason about all mentions in context, producing full cluster assignments instead of pairwise links or QA-style answers, and has been shown to elicit more coherent document-wide clusters than QA templates. Therefore we only report results for the *document template* setting.

4.4. Results

Table 2 and Table 3 compare cluster- and document-level performance across `GPT-4.1`, `Qwen-2.5-32B-Instruct`, and `Maverick`. Overall, `GPT-4.1` is the strongest model. At the document level, it achieves a CoNLL score of 0.751 overall and 0.809 on the LITERAL-only subset, showing consistent advantages across all cohorts. In contrast, `Qwen 2.5` performs substantially worse, with low precision and recall across categories, while `Maverick` achieves competitive performance.

At the cluster level, `GPT-4.1` achieves a B^3 F1 of 0.725 and a LEA of 0.725 on LITERAL clusters. More importantly, on MET_IN clusters, where metonymic readings should be integrated with their literal counterparts, it reaches a B^3 F1 of 0.647 and a LEA of 0.597, indicating a stronger ability to resolve sense shifts by placing metonymic mentions inside the correct entity cluster. In comparison, `Maverick` achieves a B^3 F1 of 0.523 and a LEA of 0.489 on MET_IN, trailing behind `GPT-4.1`.

Interestingly, `Maverick` reports higher scores on the MET_OUT subset, with a B^3 F1 of 0.563 and a LEA of 0.512, compared to 0.514 and 0.458, respectively, for `GPT-4.1`. However, this apparent improvement does not necessarily reflect a better understanding of metonymy. Because the original OntoNotes and CoNLL annotations largely excluded metonymic mentions from entity clusters, the OntoNotes-trained `Maverick` model has effectively learned to follow those conventions—often merging figurative mentions into canonical clusters even when the gold standard in our metonymy-augmented corpus expects separation. Thus, its relatively higher performance on MET_OUT is likely an artifact of alignment with the original annotation bias rather than genuine metonymy discrimination. By contrast, `GPT-4.1` appears to better distinguish when figurative mentions should or should not be coreferent, producing lower but more semantically grounded scores on this subset.

Document-level results reinforce these tendencies. For `GPT-4.1`, the CoNLL score drops from 0.809 on LITERAL-only documents to 0.735 in documents containing MET_IN mentions and further to 0.681 in those containing MET_OUT. Although performance declines in the presence of metonymy, the degradation is less severe for MET_IN, suggesting that integrating metonymic readings is comparatively easier than separating them. `Maverick`, by contrast, experiences a sharper decline—from 0.647 overall to 0.583 on MET_IN documents—while maintaining 0.672 on MET_OUT. This stability again likely reflects its inherited tendency from OntoNotes to conflate metonymic and literal references within the same cluster. Finally, `Qwen 2.5` shows the largest decline across all subsets, confirming its limited robustness to referential sense variation.

Because cluster-level metrics such as B^3 and LEA are sensitive to the number of mentions within each cluster, we further group clusters into size buckets, i.e., 2–3, 4–7, and 8+ mentions, to control for this confound. Larger clusters generally yield lower scores, as they span more mentions and discourse distance, creating more opportunities for partial matches or missed links. In contrast, smaller clusters are structurally easier and can inflate macro-averaged performance. By comparing results within these buckets in Table 4, we confirm that the relative performance differences across categories (LITERAL, METONYMIC-IN, METONYMIC-OUT) remain consistent, indicating that our findings

Category	#Cls	GPT-4.1				Qwen 2.5				Maverick			
		B^3 P	B^3 R	B^3 F1	LEA	B^3 P	B^3 R	B^3 F1	LEA	B^3 P	B^3 R	B^3 F1	LEA
LITERAL	132	0.720	0.784	**0.725**	**0.725**	0.475	0.431	0.450	0.463	0.628	0.620	0.609	0.568
MET_IN	24	0.670	0.658	**0.647**	**0.597**	0.429	0.418	0.419	0.408	0.505	0.521	0.523	0.489
MET_OUT	48	0.523	0.509	0.514	0.458	0.373	0.354	0.358	0.352	0.608	0.552	**0.563**	**0.512**

Table 2: Cluster-level B^3 and LEA by cluster category on CoNLL-Coref-Met.

Cohort	GPT-4.1				Qwen 2.5				Maverick			
	MUC	B^3	CEAF	CoNLL	MUC	B^3	CEAF	CoNLL	MUC	B^3	CEAF	CoNLL
All docs	0.810	0.752	0.731	**0.751**	0.348	0.302	0.290	0.313	0.720	0.621	0.621	0.647
Docs: LITERAL-only	0.852	0.804	0.782	**0.809**	0.395	0.343	0.338	0.359	0.772	0.682	0.690	0.715
Docs: has MET-IN	0.774	0.693	0.690	**0.735**	0.332	0.277	0.268	0.293	0.640	0.594	0.568	0.583
Docs: has MET-OUT	0.728	0.685	0.664	**0.681**	0.298	0.268	0.254	0.274	0.729	0.643	0.651	0.672

Table 3: Document-level results by document cohort on CoNLL-Coref-Met.

Category	Bucket	#Cls	B^3	LEA
LITERAL	2–3	88	0.731	0.742
	4–7	29	0.720	0.689
	8+	15	0.698	0.694
MET_IN	2–3	17	0.705	0.653
	4–7	5	0.626	0.613
	8+	2	0.654	0.622
MET_OUT	2–3	16	0.524	0.483
	4–7	19	0.547	0.426
	8+	13	0.495	0.414

Table 4: Cluster-level B^3 F1 and LEA scores across cluster size buckets.

are not driven by cluster size bias but reflect genuine variation in how models handle metonymic reference.

4.5. Qualitative Analysis

To better understand the failure modes behind the aggregate trends, we examined metonymic mentions where `GPT-4.1` and `Maverick` erred categorizing by the metonymy subtypes from Markert and Nissim (2007). Across the development set we observe a small set of recurrent phenomena that systematically elicit mistakes: (i) ORG-FOR-FACILITY and ORG-FOR-PRODUCT alternations for media/entertainment brands; (ii) ORG-FOR-MEMBERS alternations where an institution name stands for its agents; and (iii) PLACE-FOR-PEOPLE alternations in which country names stand for their governments or populations. Below we illustrate each with representative errors.

Org-for-facility / Org-for-product. Names such as *Disney*, *ABC*, and *CNN* frequently alternate between the legal entity, a physical site, and a broadcast/program. In sentence (3), *[Disney]* denotes the *park/facility* but is pulled by the model toward the *company* cluster, consistent with a lexical prior that prefers the corporate sense over local physical realizations. In sentence (4), *[ABC News]* refers to an on-air segment yet is merged into the organization cluster. These errors tend to occur when local syntactic cues are weak and the governing predicate is not leveraged to disambiguate sense.

(3) The subway to **[Disney]** has already been constructed.

(4) He was speaking live on **[ABC News]**.

Org-for-members. Institution names often stand for their members, e.g., journalists, agents, officials. In sentence (5) and (6), *the FBI* and *CNN* are used in ORG-FOR-MEMBERS senses. Models frequently fail to link these to the correct entity cluster, suggesting insufficient use of event frames and selectional preferences: predicates like *confirm*, *announce*, and *allege* favor an animate/agentive reading and should trigger a metonymic link to the institutional entity.

(5) The authorities, including **[the FBI]**, ...

(6) Later, **[CNN]** confirmed ...

Place-for-people. Country names routinely alternate between the geopolitical territory and its government or populace. In Example 7 and 8, the metonymic readings are expected to be disambiguated from the literal place cluster, but the model collapses them. We find such collapses are amplified by topical cohesion: repeated country names within a paragraph bias the model toward a single coarse cluster, overwhelming finer-grained sense distinctions signaled by predicates like *accuse*, *negotiate*, or by human-collective nominals like *the Israelis*.

Cluster-level metrics				
Category	**Original**		**Reclustered**	
	B^3	LEA	B^3	LEA
LITERAL	0.725	0.725	**0.731**	**0.749**
MET_IN	0.647	0.597	**0.674**	**0.627**
Document-level CoNLL F1				
Cohort	**Original**		**Reclustered**	
	CoNLL		CoNLL	
All docs	0.751		**0.758**	
Docs: has MET_IN	0.735		**0.765**	

Table 5: Results for `GPT-4.1` on CoNLL-Coref-Met-Reclustered (Original) and CoNLL-Coref-Met-Reclustered (Reclustered). After reclustering, MET_OUT is not applicable.

Cluster-level metrics				
Category	**Direct**		**Metonymy-aware**	
	B^3 F1	LEA	B^3 F1	LEA
LITERAL	**0.731**	0.749	0.729	**0.754**
MET_IN	0.674	0.627	**0.680**	**0.701**
Document-level CoNLL F1				
Cohort	**Direct**		**Metonymy-aware**	
	CoNLL		CoNLL	
All docs	0.758		**0.778***	
Docs: has MET_IN	0.765		**0.793**	

Table 6: Results of direct and metonymy-aware coreference resolution for `GPT-4.1`. * indicates statistically significant improvement over the direct baseline ($p < 0.05$) under paired bootstrap resampling with 10,000 samples at the document level.

(7) ...persuade **[North Korea]** to give up its missile program ...

(8) Mubarak ...accused **[Israel]** of ...

The qualitative patterns align with the quantitative results above. `GPT-4.1`'s relative strength on MET_IN suggests it more often *integrates* ORG-FOR-MEMBERS, ORG-FOR-PRODUCT, and ORG-FOR-FACILITY readings into the canonical cluster when appropriate, likely due to better exploitation of event semantics and discourse cues; yet it can still over-integrate in MET_OUT contexts. `Maverick`'s tendency to under-attach MET_IN mentions but avoid some MET_OUT merges reflects a clustering bias toward preserving surface distinctions without sufficiently modeling predicate-licensed metonymy.

4.6. Reclustering Results

To test whether metonymy-aware normalization reduces evaluation penalties, we compare the model predictions against (i) the original gold and (ii) the reclustered gold. This isolates the effect of gold normalization without conflating it with changes in model outputs. We report the results in Table 5.

After reclustering, the MET_OUT category disappears as intended. By folding MET_OUT mentions into their referent clusters yields small but consistent improvements for `GPT-4.1` at both cluster and document levels, and provides a cleaner testbed for assessing whether systems *integrate* metonymic readings into the correct entities rather than segregating them.

5. Metonymy-Aware Coreference Resolution

We further explore whether explicit metonymy cues can guide coreference resolution toward intended semantic reference rather than surface identity. Our approach is a two-stage pipeline that first resolves whether each mention in a document is used literally or metonymically and, second, performs identity coreference at the document level while using those metonymy cues. The method is designed to answer a simple question: when figurative language is present, can an LLM cluster mentions by their intended referent rather than by surface form?

Stage 1: Metonymy resolution. Given the marked document, the LLM first decides, for every mention, whether its use in context is LITERAL or METONYMIC. If metonymic, the model assigns a subtype from the SemEval 2007 Task 8 typology (Markert and Nissim, 2007). This stage yields a metonymy layer over the document. The layer serves as an contextual cue that can be projected onto any existing chains for analysis.

Stage 2: Metonymy-aware coreference. The same document is then clustered using the LLM document template from Le and Ritter (2023) in which the Stage A metonymy labels are presented as cues. The metonymy labels and subtypes are introduced to help the model interpret each mention's facet.

We report the metonymy-aware resolution results for `GPT-4.1` on CoNLL-Coref-Met-Reclustered alongside the results of direct resolution in Table 6. We assess statistical significance using paired bootstrap resampling over documents with 10,000 samples. The comparison shows that incorporating metonymy cues yields consistent numerical gains over direct coreference resolution without such information. At the cluster level, the metonymy-aware setting improves performance on MET_IN clusters, raising B^3 F1 from 0.674 to 0.680 and LEA from 0.627 to 0.701. This pattern suggests that explicit modeling of metonymic links helps the model attach figurative mentions to their intended entity clusters.

On LITERAL clusters, performance remains essentially stable, with B^3 F1 changing only from 0.731 to 0.729 and LEA from 0.749 to 0.754, indicating that the added metonymy cues do not materially harm standard coreference decisions.

At the document level, the overall CoNLL F1 improves from 0.758 to 0.778, and this gain is statistically significant under paired bootstrap resampling ($p < 0.05$). On documents containing MET_IN mentions, CoNLL also increases from 0.765 to 0.793, although we do not mark this subset gain as statistically significant. Taken together, these results suggest that metonymy-aware prompting can improve document-level coherence while preserving performance on literal cases, helping the model better reconcile literal and figurative uses of the same discourse referent.

6. Conclusion

Metonymy challenges the assumption that identical surface forms always denote identical entities, revealing a blind spot in how coreference systems interpret meaning. In this work, we introduced CoNLL-Coref-Met, a metonymy-aware extension of CoNLL-2012 that enables systematic evaluation of how figurative reference affects clustering behavior. Our analyses show that both neural and LLM-based resolvers degrade significantly when metonymy is present, often merging distinct referents or fragmenting single entities across surface facets.

By reclustering mentions based on metonymic interpretation and introducing a metonymy-aware resolution pipeline, we demonstrated that incorporating explicit cues about literal and figurative usage improves model coherence without sacrificing precision. These results highlight the need to move beyond identity-only formulations of coreference and to integrate semantic interpretation directly into resolution pipelines. We hope that CoNLL-Coref-Met will serve as a foundation for future research on bridging literal and figurative reference in discourse-level understanding.

7. Limitations

Scope of annotation. Our metonymy layer targets only Organization and Location mentions and is restricted to the English CoNLL-2012 *development* portion. This design isolates a common and consequential subset of metonymic phenomena but limits coverage: other entity types (e.g., Person, Facility, Product) and other genres/languages may exhibit different patterns. Consequently, external validity beyond newswire-style text and beyond org/place metonymy is not guaranteed.

Dataset size and statistical power. The development-only scope (94 documents; 204 clusters) constrains the number of metonymy-bearing cases, especially when further stratified by subtype and cluster size. Some observed differences may be underpowered or sensitive to modest distributional shifts. In addition, the relatively small dataset size prevents meaningful model training or fine-tuning, constraining our analysis to evaluation and comparison rather than supervised adaptation.

Annotation bias and LLM assistance. Metonymy flags were produced via an LLM-assisted, human-in-the-loop workflow. Although all labels were verified, residual model-driven biases and anchoring effects may persist. In particular, borderline cases (e.g., media brands oscillating between ORG-FOR-EVENT and ORG-FOR-PRODUCT) remain difficult, and subtle discourse cues can be missed in long contexts. Future versions should incorporate double-blind adjudication and a larger, independently curated pilot for quality calibration.

8. Acknowledgements

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grants DRL 2019805 and DRL 2454151. The opinions expressed are those of the authors and do not represent the views of the NSF.

We are grateful to the anonymous reviewers for their insightful feedback; any remaining errors are, of course, our own.

9. Bibliographical References

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amit Bagga. 1998. Algorithms for scoring coreference chains. In *Proc. Linguistic Coreference Workshop at the first Conf. on Language Resources and Evaluation (LREC), Granada, Spain, May 1998*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Rakesh Chada. 2019. Gendered pronoun resolution using BERT and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133, Florence, Italy. Association for Computational Linguistics.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4833–4843, Singapore. Association for Computational Linguistics.

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gilles Fauconnier. 1994. 1985. mental spaces.

Tim Finin, Zareen Syed, James Mayfield, Paul McNamee, Christine D Piatko, et al. 2009. Using wikitology for cross-document entity coreference resolution. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*, pages 29–35.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Saptarshi Ghosh and Tianyu Jiang. 2025. ConMeC: A dataset for metonymy resolution with common nouns. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6493–6509, Albuquerque, New Mexico. Association for Computational Linguistics.

Elisabetta Ježek. 2016. *The lexicon: An introduction*. Oxford university press.

Mark Johnson and George Lakoff. 1980. *Metaphors we live by*, volume 1. University of Chicago press Chicago.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Nghia T Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *arXiv preprint arXiv:2305.14489*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Haonan Li, Maria Vasardani, Martin Tomko, and Timothy Baldwin. 2020. Target word masking for location metonymy resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3696–3707, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Katja Markert and Malvina Nissim. 2002. Metonymy resolution as a classification task. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 204–213. Association for Computational Linguistics.

Katja Markert and Malvina Nissim. 2007. SemEval-2007 task 08: Metonymy resolution at SemEval-2007. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 36–41, Prague, Czech Republic. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 56–63, Sapporo, Japan. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

James Pustejovsky. 1995. *The generative lexicon*. MIT press.

James Pustejovsky and Anna Rumshisky. 2009. SemEval-2010 task 7: Argument selection and coercion. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 88–93, Boulder, Colorado. Association for Computational Linguistics.

Emma Romani and Elisabetta Ježek. 2020. Tracing metonymic relations in t-pas: An annotation exercise on a corpus-based resource for italian. *Computational Linguistics CLiC-it 2020*, page 373.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Sameer Singh, Amarnag Subramanya, Fernando Pereira, and Andrew McCallum. 2011. Large-scale cross-document coreference using distributed inference and hierarchical models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 793–803, Portland, Oregon, USA. Association for Computational Linguistics.

Natalia Skachkova. 2024. Multilingual coreference resolution as text generation. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 114–122, Miami. Association for Computational Linguistics.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association*, 19(5):786–791.

Hao Wang, Siyuan Du, Xiangyu Zheng, and Lingyi Meng. 2023. An empirical study of incorporating syntactic constraints into bert-based location metonymy resolution. *Natural Language Engineering*, 29(3):669–692.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Patrick Xia, João Sedoc, and Benjamin Van Durme. 2020. Incremental neural coreference resolution in constant memory. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8617–8624, Online. Association for Computational Linguistics.

Bingyang Ye, Jingxuan Tu, Elisabetta Jezek, and James Pustejovsky. 2022. Interpreting logical

metonymy through dense paraphrasing. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.