

# Can LLMs Detect Ambiguous Plural Reference? An Analysis of Split-Antecedent and Mereological Reference

Anh Dang<sup>1</sup>, Rick Nouwen<sup>1</sup>, Massimo Poesio<sup>1,2</sup>

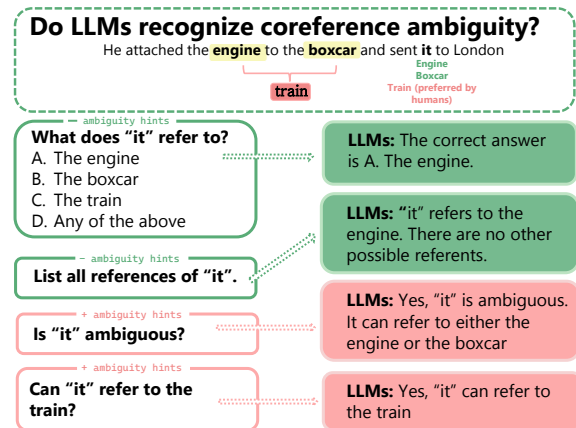
<sup>1</sup>Utrecht University, <sup>2</sup>Queen Mary University of London  
{t.t.a.dang, R.W.F.Nouwen, m.poesio}@uu.nl

## Abstract

Our goal is to study how LLMs represent and interpret plural reference in ambiguous and unambiguous contexts. We ask the following research questions: (1) Do LLMs exhibit human-like preferences in representing plural reference? (2) Are LLMs able to detect ambiguity in plural anaphoric expressions and identify possible referents? To address these questions, we design a set of experiments, examining pronoun production using next-token prediction tasks, pronoun interpretation, and ambiguity detection using different prompting strategies. We then assess how comparable LLMs are to humans in formulating and interpreting plural reference. We find that LLMs are sometimes aware of possible referents of ambiguous pronouns. However, they do not always follow human reference when choosing between interpretations, especially when the possible interpretation is not explicitly mentioned. In addition, they struggle to identify ambiguity without direct instruction. Our findings also reveal inconsistencies in the results across different types of experiments.

## 1 Introduction

Large language models (LLMs) have been shown to reach human-like competence in a wide range of natural language tasks (Touvron et al., 2023; Brown et al., 2020). Communication between humans and AI systems is becoming more ubiquitous, given their remarkable language understanding and generation capacity. An important factor in effective communication is the ability to handle ambiguity in language. Although humans are very good at resolving ambiguity based on context and common sense knowledge, LLMs are designed with a strong focus on accuracy. In fact, they have been shown to struggle with capturing language ambiguity (Liu et al., 2023; Wildenburg et al., 2024; Stengel-Eskin et al., 2023; Karamolegkou et al., 2025). They are



**Figure 1:** An illustration of the setup and results of our prompting experiments about how the extent to which LLMs can detect ambiguity in anaphoric expressions. Psycholinguistic studies show that in the ambiguous sentence above, humans prefer using *it* to refer to combination of the *engine* and the *boxcar* or to one of the objects. We found that LLMs’ responses vary depending on how much information about ambiguity is present in the prompt.

often unable to detect or generate ambiguous expressions. Since ambiguity is always present in communication, this inadequacy can lead to misunderstanding and the inability of LLMs to fully comprehend the use of language in context.

In this work, we focus on the ambiguity in coreference resolution. A lot of work has been done on building coreference resolution systems (Yu et al., 2020; Lee et al., 2017; Bohnet et al., 2023). More recent efforts to evaluate LLMs have shown that they can resolve coreference to some extent, yet it is still a challenging task (Brown et al., 2020; Pradhan et al., 2012; Gan et al., 2024). In actual communication, anaphoric expressions can sometimes be ambiguous. This means that it is not always clear to which referent the anaphoric expression refers. For example, when we say *Jane looked at Mary and she smiled*, the pronoun *she* possibly refers to Jane

or Mary. Without further context, it is impossible to disambiguate the pronoun.

Some research has been conducted on how LLMs process anaphoric expressions. However, most of the work focuses on singular reference and specific cases of ambiguity such as gender-ambiguous pronouns (Webster et al., 2018; Emami et al., 2019). No research has been done specifically on the processing of plural reference. In this paper, our aim is to study how LLMs conceptualize plural reference in both ambiguous and unambiguous cases. We find plural reference particularly interesting because it is highly dependent on how humans conceptualize plurality. In psycholinguistic research, formulating plural expressions is a complex process because they introduce ambiguity to the discourse (Cokal et al., 2023; Koh and Clifton Jr, 2002; Moxey et al., 2012, 2004; Asher and Wang, 2003; Patson, 2014). Our goal is to study **how LLMs produce, interpret, and represent plural reference in unambiguous and ambiguous contexts.**

We design a set of experiments, including next-token prediction and prompting, to compare how comparable LLMs are to humans in producing and interpreting ambiguous and unambiguous plural reference. We ask the following research questions.

**RQ1:** *Do LLMs exhibit human-like preferences when representing plural reference?*

**RQ2:** *Are LLMs able to detect ambiguity in plural anaphoric expressions and identify possible referents?*

Our main findings are as follows: (1) LLMs follow human preference when choosing between interpretations in ambiguous cases where the possible referents are directly mentioned and struggle to do so when they are not present in the input; (2) LLMs fail to identify ambiguity without direct instruction.

## 2 Background

There is a body of work in psycholinguistics that shows that plural reference is more difficult to process and interpret than singular reference because the antecedents of plural anaphoric expressions are usually underspecified (Patson, 2014), leaving more possibilities to interpret them. For example, in the pair of sentences below, the pronoun *they* is more likely to refer to three people in 1(a) than 1(b). In 1(b), *they* may also refer to the group of

Tom and Dick.

- (1) a. Tom, Dick, and Harry went to a restaurant last night. They ate a pizza.
- b. Tom and Dick went to a restaurant with Harry last night. They ate a pizza.

Here, we focus specifically on the cases where there is an interaction between singular and plural reference. These are cases where singular and plural expressions are both probable. Moxey et al. (2012) find that the degree of overlap in roles between objects affects whether they are grouped as a plural entity. For example, in sentences 2(a) and 2(b), *waitress* is more related to *waiter* than *customer* in that they are more likely to perform the same action. As such, *they* is more likely to be the continuation of 2(a) than 2(b).

- (2) a. The waiter and the waitress were in the restaurant. \_\_ (He/She/They)
- b. The waiter and the customer were in the restaurant. \_\_ (He/She/They)

More recently, Cokal et al. (2023) investigate how humans interpret singular and plural pronouns in contexts where there is a whole-part relation between entities. When two singular objects are introduced to the discourse, and they are combined into a unified object, the new unified object is known as a *mereological* object (Poesio et al., 2006). It has been found that when more than one object is present, a plural expression is preferred and the use of a singular expression can cause processing difficulty (Moxey et al., 2004; Sanford and Moxey, 1995; Gordon et al., 1999). However, there are cases where a singular reference may become salient as a new entity is created by combining the two antecedents.

- (3) a. The engineer attached the *engine* to the *boxcar* and sent \_\_ (it/them)
- b. The engineer detached the *engine* from the *boxcar* and sent \_\_ (it/them)

Consider the two sentences above. In 3(a), through the verb *attached*, the engine and the boxcar are combined, introducing a new singular object to the discourse. This new object is a *train*, though it is never explicitly mentioned. Now, a singular or plural reading is both appropriate. The pronoun following the verb *sent* can be *it* or *them*, with *it* referring to one of the parts (that is, the engine or

the boxcar) or the new object (that is, the train). In 3(b), however, the verb *detached* prevents the formation of the mereological entity, as such, when *it* is used, it can only refer to one of the objects. *Them* is used to refer to the group of two objects. Cokal et al. (2023) found that humans prefer using *it* more in 3(a) than 3(b). Moreover, they are more likely to consider the mereological object rather than the constituents to be the referent of *it*. This suggests that when entities are joined, a new unified *group* entity is created, making *it* a preferred antecedent for singular pronouns. In the following sections, we study how LLMs produce and interpret pronouns in these ambiguous cases.

### 3 Ambiguity in LLMs

Ambiguity is when an expression can be interpreted in different ways in a context. The use of everyday language is full of ambiguity, and it is shown that ambiguity indeed facilitates effective communication by minimizing effort (Piantadosi et al., 2012). With LLMs, recognizing ambiguity can be challenging, as they are not trained for this purpose and do not have access to pragmatic contexts.

An emerging body of research has been done on how LLMs handle ambiguity in input. Stengel-Eskin et al. (2023) focus on semantic parsing. They prompt LLMs with ambiguous sentences with two possible interpretations and ask them to respond with the logical parses that correspond to the interpretations. They found that LLMs are unable to capture the possible interpretations in zero-shot contexts but perform much better in the few-shot contexts. Sedova et al. (2024) look at entity ambiguity. They show that while LLMs are aware of the possible readings, they often struggle with explicitly verbalizing the ambiguity. Liu et al. (2023) use a prompting-based approach to test the ability of LLMs in disambiguating, recognizing, and modeling different types of linguistic ambiguity. They found that even GPT-4 still struggles with capturing ambiguity.

In the domain of coreference resolution, some studies have considered coreference to be one of the main sources of ambiguity (Wildenburg et al., 2024; Kamath et al., 2024; Liu et al., 2023; Gan et al., 2024). However, the types of ambiguity they cover are quite limited (Wildenburg et al., 2024; Kamath et al., 2024; Liu et al., 2023) or strongly focus on unambiguous reference (Gan et al., 2024). The instances related to coreference account for

a small number in their datasets. As such, in this work, we aim for a more extensive investigation of different types of plural reference ambiguity.

## 4 Methods

We use two types of experiments to study how LLMs process ambiguous plural references. For both experiments, we test a number of recent autoregressive decoder-only LLMs, namely Qwen2.5-7B (Team, 2024), Aya-Expansive-8B (Dang et al., 2024) and Llama-3.1-8B (Touvron et al., 2023). We evaluate the ability of LLMs to process ambiguous plural anaphoric expressions along the following dimensions. These capabilities are tested using a combination of prompting and next-token prediction tasks.

**Production** is the ability to predict the correct pronoun for the referents. We test whether LLMs are aware of the pronouns for all possible antecedents. For example, in mereological contexts, are they able to identify that both singular (mereological object) and plural pronouns (constituents) are appropriate?

**Interpretation** means being able to directly specify the possible antecedents of the pronoun. Several prompting experiments will be used to evaluate this ability.

**Detection** refers to the ability to identify whether the pronoun can have multiple possible antecedents. This ability can be demonstrated through the extent to which LLMs can provide meta-linguistic judgment of the ambiguity.

## 5 Production

### 5.1 Simple Plural References

We first look at the case of simple plural reference as a baseline experiment, where the referents of the pronouns are completely unambiguous. When the direct object is singular, the pronoun referring to that object must also be singular. In the experiment, we asked the LLMs to predict the pronoun for both singular and plural referents.

**Dataset** We use GPT-4o to generate the dataset<sup>1</sup>. It consisted of 74 pairs of sentence prefixes. After the generation, we manually inspect each sentence

---

<sup>1</sup>The GitHub repository consisting of the datasets and code is available at <https://github.com/dangthithaoanh/ambiguous-plural-reference>

Model	Condition	$P_{sing}$	$P_{plur}$	Diff.
Aya	Singular	.336	.009	.326
	Plural	.014	.484	-.470
Llama	Singular	.352	.009	.343
	Plural	.012	.467	-.455
Qwen	Singular	.576	.007	.568
	Plural	.021	.478	-.456

**Table 1:** Results of Probability Prediction Task for Unambiguous Reference. Diff. is the probability difference between singular and plural pronouns ( $P_{it} - P_{them}$ ), averaged across examples.

pair to ensure that they meet the following requirements: (1) The direct object must be a countable object and (2) the second verb must be plausible with both the singular and plural versions of the direct object. Each pair differs in the number of the direct object, causing a difference in the choice of pronoun following the second verb. For each sentence prefix, there is only one correct pronoun. See the pair of sentences below for an example.

- (4) a. The Michelin chef prepared the dish and served \_\_\_\_  
b. The Michelin chef prepared the dishes and served \_\_\_\_

**Results** Table 1 shows the mean probability of singular and plural pronouns and the difference between them. As expected, we found that LLMs produce much higher probabilities for the correct pronouns for the sentence prefixes. This shows that, in simple unambiguous contexts, LLMs are fully capable of producing correct pronouns for the antecedents.

## 5.2 Split Antecedents Reference

Moxey et al. (2012) find that the degree of overlap between roles assigned to characters predicts the formation of a plural reference object. If LLMs exhibit a preference similar to that of humans, we would expect that the probabilities of the plural pronoun in 2(a) are higher than in 2(b), as shown in Moxey et al. (2012). Following the study, we include another condition in which the two characters are linked by the conjunction *with* instead of *and*. There are studies showing that the conjunction *with* may decrease the preference for using plural pronouns (Albrecht and Clifton, 1998; Sanford and Lockhart, 1990). The pair of sentences below shows the difference between the *and* and *with* conditions.

- (5) a. The waiter and the waitress were in the restaurant. \_\_\_\_  
b. The waiter was in the restaurant with the waitress. \_\_\_\_

**Dataset** The dataset is adopted from the materials in Moxey et al. (2012). Each condition consists of 24 sentence prefixes. The conditions include one with a high overlap between roles or spatial status. In addition to the overlap of the roles, there are also two other conditions where the conjunction linking the two roles is different, namely *and* and *with*. Using a 2x2 setting, this results in 4 conditions per example. See Table 5 for an example of all conditions.

**Metric** For each prefix, we extract and compare the probability difference between singular and plural pronouns. Possible singular pronouns include *he* and *she* ( $P_{he/she}$ ). We consider the pronoun with the highest probability. Plural pronoun is *they* ( $P_{they}$ ).

Model	Condition	$P_{sing}$	$P_{plur}$	Diff.
Aya	High-And	.015	.163	-.148
	High-With	.047	.145	-.097
	Low-And	.016	.100	-.084
	Low-With	.047	.102	-.054
Llama	High-And	.013	.129	-.116
	High-With	.052	.153	-.101
	Low-And	.014	.065	-.051
	Low-With	.065	.116	-.051
Qwen	High-And	.003	.151	-.147
	High-With	.022	.125	-.103
	Low-And	.003	.075	-.072
	Low-With	.039	.088	-.048

**Table 2:** Results of Probability Prediction Task for Split-antecedent Plural References. The *High* and *Low* conditions represent high or low overlap between roles. *And* and *With* are the conjunction linking the two roles. Diff. is the probability difference between the singular and plural pronoun ( $P_{sing} - P_{plur}$ ), averaged across examples.

**Results** We compare the probabilities of singular and plural continuations between conditions. Table 2 reports the mean probabilities between the targeted pronouns in all conditions for both LLMs. We run a two-way ANOVA with role overlap and conjunction as independent variables. The dependent variable is the probability difference between the singular and plural pronouns.

We find that the main effect of role overlap was significant (Aya-Expanse-8B:  $F(1, 23) = 35.04$ ,  $p < .001$ ; Qwen2.5-7B:  $F(1, 23) = 16.7$ ,  $p < .001$ ; Llama-3.1-8B:  $F(1, 23) = 17.96$ ,  $p < .001$ ).



The main effect of conjunction was also not significant for Aya-Expanse-8B ( $F(1, 23) = 1.23, p = .277$ ) but was significant for Qwen2.5-7B ( $F(1, 23) = 5.85, p = .023$ ) and Llama-3.1-8B ( $F(1, 23) = 33.37, p < .001$ ). The interaction effect between conjunction and role overlap was not significant (Aya-Expanse-8B:  $F(1, 23) = 1.23, p = .277$ ; Qwen2.5-7B:  $F(1, 23) = .44, p = .512$ ; Llama-3.1-8B:  $F(1, 23) = 1.54, p = .226$ ). This result is consistent with Moxey et al. (2012), suggesting that the degree of overlap of the roles between the antecedents affects how LLMs assign probabilities to singular and plural pronouns.

### 5.3 Mereological Reference

We report experiments on the production of singular and plural pronouns in mereological contexts, a type of ambiguous plural reference where two distinct objects also form a unified object.

**Dataset** The dataset of this task consists of 44 pairs of mereological and non-mereological sentences. Half of the dataset is adapted from the material from Cokal et al. (2023). We manually generate more pairs of sentences following the same template. Before running the main experiment, we want to ensure that the model is aware of the mereological object formed by attaching the two objects. We annotate the dataset with the names of the mereological objects and create another version of the dataset, in which the original sentence prefixes are appended with the token *the*. Each sentence in the dataset for this experiment thus has the following template: {The engineer attached the engine to the goods wagon and sent the; engine; wagon; train}

We then have the LLMs predict the next token of the mereological sentences, where the mereological objects may have a higher chance of being observed. We then look at the top 15 tokens with the highest probabilities. If the mereological object is one of those tokens, then the sentence prefix is qualified. Otherwise, we remove the prefix from the dataset. This results in a total of 44 pairs of sentences.

#### 5.3.1 Experiment 1: Pronoun Production

In this experiment, we test whether LLMs prefer singular pronouns in mereological contexts when the new objects are created. Through a next-word prediction task, we extract the probabilities of *it* and *them* in the mereological and non-mereological

Model	Type	$P_{it}$	$P_{them}$	Diff.
Aya	Mereo.	.336	.009	.326
	Non-mereo.	.484	.014	.470
Llama	Mereo.	.352	.009	.343
	Non-mereo.	.467	.012	.455
Qwen	Mereo.	.363	.005	.357
	Non-mereo.	.507	.012	.494

**Table 3:** Results of Probability Prediction Task for Mereological Reference. The Diff. column represents the mean probability difference between *it* and *them* ( $P_{it} - P_{them}$ ).

sentences. If the pronoun is *it*, it refers to the individual object (that is, engine or boxcar) or the mereological object (i.e., train).

**Metric** We compare the probability difference between *it* and *them* in the mereological ( $D_{mereo}$ ) and non-mereological ( $D_{N.mereo}$ ) sentences. The following formulas show how we calculate the difference.

$$D_{mereo} = P_{it|mereo} - P_{them|mereo}$$

$$D_{N.mereo} = P_{it|N.mereo} - P_{them|N.mereo}$$

If the difference is larger in the mereological sentences than the non-mereological sentences, then the LLMs show a higher preference for *it*, as found in Cokal et al. (2023).

**Results** Table 3 demonstrates how LLMs assign probabilities to *it* and *them* in mereological and non-mereological cases. It can be seen that LLMs prefer using *it* in both contexts. We run paired-sample *t*-tests, comparing  $D_{mereo}$  and  $D_{N.mereo}$  across all pairs of sentences. The results show that the difference is significantly larger in non-mereological contexts (Aya-Expanse-8B:  $t = -4.78, p < .001$ ; Llama-3.1-8B:  $t = -3.88, p < .001$ ; Qwen2.5-7B:  $t = -5.73, p < .001$ ). The statistical results show the opposite of what we expected. LLMs give higher probability to the singular pronoun when processing non-mereological sentences. It should be noted that this result does not reflect whether LLMs give a higher preference to the mereological object in mereological contexts because we do not yet know what *it* refers to. As such, we run another experiment, in which we consider the probabilities of the objects themselves.

#### 5.3.2 Experiment 2: Object Production

After the first experiment, we observe that *it* has very high probabilities in both contexts. However, while the pronoun *them* clearly refers to the group of two objects, the antecedent of the pronoun *it*

can be highly ambiguous. It can refer to either the mereological object or one of the constituents. We also observe that the probability of *the* is quite high: in most cases, the next most probable token is *the*, which can be followed by the individual object or the mereological object. Thus, evidence for mereological object preference can also be demonstrated when the direct object of the verb is specified. As such, we conducted another experiment with *the* after the original prefix. Adding *the* forces the LLMs to specify the object to be referenced.

**Dataset** We use the same dataset from Experiment 1 and append *the* to the sentence prefixes. See Table 5 for a complete example of the dataset.

**Metric** We compare the probabilities of the individual objects with higher probabilities ( $P_{ind.}$ ) and the mereological object ( $P_{mereo.}$ ). See the formulas below for the detailed calculation of the metric.

$$D_{mereo} = P_{ind.|mereo} - P_{mereo.|mereo}$$

$$D_{N.mereo} = P_{ind.|N.mereo} - P_{mereo.|N.mereo}$$

If the probability difference between the mereological objects and the constituents in the mereological cases ( $D_{mereo}$ ) is higher than that in the non-mereological cases ( $D_{N.mereo}$ ), then the LLMs show a greater preference to choose the mereological objects as the referent for the pronoun *it* in mereological contexts.

Model	Type	$P_{ind.}$	$P_{mereo.}$	Diff.
Aya	Mereo.	.076	<b>.114</b>	-.038
	Non-mereo.	.099	.054	.044
Llama	Mereo.	.091	.081	.009
	Non-mereo.	.134	.048	.085
Qwen	Mereo.	.059	<b>.191</b>	-.132
	Non-mereo.	.206	.045	.160

**Table 4:** Results of Probability Prediction Task for Object Production Experiment. Diff. is the mean probability difference between  $P_{ind.}$  and  $P_{mereo.}$ .

**Results** In Table 4, we display the mean probability of individual objects, the mereological objects, as well as the difference between them. For Aya-Expanse-8B and Qwen2.5-7B, the mean probability of the mereological object is higher than its constituents. Paired-sample *t*-tests show that  $D_{mereo}$  is significantly larger than  $D_{N.mereo}$  (Aya-Expanse-8B:  $t = -2.845$ ,  $p = .006$ ; Llama-3.1-8B:  $t = -2.669$ ,  $p = .01$ ; Qwen2.5-7B:  $t = -5.825$ ,  $p < .001$ ). This result suggests that when directly choosing an object for the second verb (e.g.

*sent*), the LLMs do prefer the joined object more in the mereological contexts.

## 6 Interpretation

While Section 5 aims at testing whether LLMs are able to produce correct pronouns in unambiguous and ambiguous contexts, in this Section we look at how LLMs interpret these pronouns. Do they interpret them differently in ambiguous contexts? Do they notice possible ambiguity? To do this, we use various prompting techniques to study the extent to which LLMs can detect ambiguity and generate possible referents for the pronouns.

Previous studies have found that while LLMs may be aware of the potential ambiguity in input, they are sometimes unable to transform this knowledge into a textual explanation (Stengel-Eskin et al., 2023; Karamolegkou et al., 2025). Taking into account such a possibility, we use two prompting directions. In Prompt 1 (**P1**), we ask LLMs to provide all possible referents for a pronoun. In Prompt 2 (**P2**), LLMs have to choose the referent for the pronoun among those in a supplied list. Some of the listed referents suggest ambiguity, to see to what extent the LLMs can make use of them to detect ambiguity. In Prompt 3 (**P3**), we ask LLMs if the mereological object can be the referent. Finally, we ask if the LLMs think that the pronoun *it* is ambiguous (**P4**). All of our prompts are few-shot. See Appendix A for details about our prompt designs and how we code LLMs’ responses.

For baseline comparison, we also prompt LLMs on the simple reference dataset. The experiments are run on two LLMs, namely Aya-Expanse-8B and Qwen2.5-7B.

### 6.1 Reference Identification

**Unambiguous Reference** As a baseline, we first test whether LLMs can choose the correct antecedent for simple unambiguous reference using the dataset described in Section 5.1. The results are not surprising. The LLMs reach 100% accuracy. This shows that in unambiguous contexts, LLMs are fully able to choose the correct antecedents for pronouns.

**Ambiguous Reference** We report the results of **P1** and **P2** for the mereological dataset. When being asked to list all possible referents for *it*, the LLMs show somewhat different patterns. Aya-Expanse-8B consistently chooses one of the

Reference Type	Sentence prefix	Expected pronoun	Metrics
Simple Singular/Plural	The chef prepared the dish and served __	it /them	$P_{it} - P_{them}$
	The chef prepared the dishes and served __	it /them	
Split-Antecedent Moxey et al. (2004, 2012)	The waiter and the waitress were in the restaurant. __	he / she / they	$P_{he/she} - P_{them}$
	The waiter was in the restaurant with the waitress. __		
	The waiter and the customer were in the restaurant. __		
	The waiter was in the restaurant with the customer. __		
Mereological Cokal et al. (2023)	The engineer attached the engine to the goods wagon and sent __	it / them	$D_{mereo} - D_{N.mereo}$
	The engineer detached the engine from the goods wagon and sent __		
	The engineer attached the engine to the goods wagon and sent the __	mereo. obj. / ind obj.	$D_{mereo.} - D_{N.mereo}$
	The engineer detached the engine from the goods wagon and sent the __		

**Table 5:** Examples of the datasets used in all experiments. The types with expected pronouns in blue mean that there can only be one possible pronoun and it is unambiguous. Those in pink mean that multiple pronouns are possible.

constituents as the only referent in both mereological and non-mereological contexts. Qwen2.5-7B is better at recognizing that the pronoun is ambiguous and that both objects can be the referent (50%). However, both rarely mention the mereological object as a possible referent (11.3%).

Figure 2 (top, left) shows the percentage of each answer choice in mereological and non-mereological contexts. When LLMs are forced to choose an antecedent for pronoun *it*, they generally prefer to link *it* to the first object. Notably, for Qwen2.5-7B, the number of times the LLMs choose the first object is much higher in the non-mereological cases than in the mereological cases. However, Aya-Expanses-8B gives nearly equal preference for the first object and the mereological object. The preference for the second object is quite similar in both cases. As a result, the percentage of choosing the mereological object is higher in the mereological cases (Aya-Expanses-8B: 37.3%, Qwen2.5-7B: 22%) than the non-mereological cases (22.8%). This shows that LLMs give higher preference to the mereological object in mereological contexts. However, they seem to be more likely to link the pronoun *it* to the constituent in both cases.

Another answer choice that we are also interested in is *D. any of the above objects*. If the LLMs choose *D*, they think that all objects are equally likely to be the referent for the pronoun *it*. In fact, if the models understand that *it* is ambiguous, this option has to be the most-chosen one. The low percentage of this option shows that, when being asked, the LLMs tend to specify an object as a referent instead of accepting the ambiguity. This shows that when the LLMs are instructed towards choosing a single referent, they often fail to identify the ambiguity that is present.

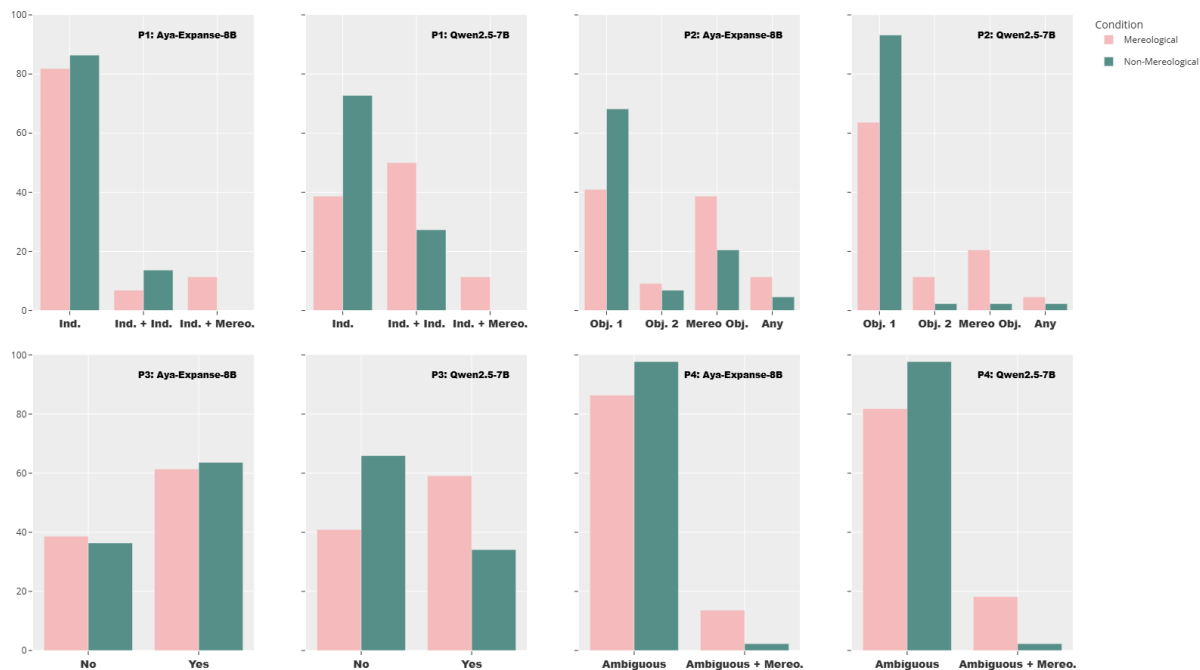
A surprising result is that Aya-Expanses-8B chooses *it* as referring to the mereological object quite often in the non-mereological prompts (21.3%), where the mereological object is not available as a possible antecedent. This can be an effect of word frequency, such that the mereological object is more frequent or more likely to be preceded by the main verb.

## 6.2 Ambiguity Detection

We now examine the results of **P3** and **P4** where we test the extent to which LLMs can detect ambiguity given direct mention. Recall that in **P4**, we ask the LLMs if *it* is ambiguous. We also run this experiment on unambiguous sentences. We find that in 97.7% of the cases, they correctly identify the pronoun as unambiguous.

Figure 2 (bottom, right) shows the percentage of *Yes* and *No* responses. We found that LLMs consistently identify the pronoun *it* as an ambiguous one (100%) in both mereological and non-mereological cases. This is rather surprising given that in the **P2** they almost never choose the option that indicates ambiguity (i.e. any of the above objects). In addition, in the explanations of LLMs, they also generate possible antecedents of *it*. In 13.6% (Aya-Expanses-8B) and 18.2% (Qwen2.5-7B) of the mereological prompts, they include the mereological object as a possible antecedent, compared to 2.3% in the non-mereological prompts.

In **P3**, we ask the LLMs if the mereological object can be the referent of *it*. The percentage of *Yes* in the mereological cases is moderate for both LLMs (Aya-Expanses-8B: 61.4%, Qwen2.5-7B: 59.1%).



**Figure 2:** Results of prompting experiments with four different prompts (P1, P2, P3, and P4). The  $x$ -axis displays different types of LLMs’ responses. In **P1**, *Ind.* is the one of the constituent. *Ind. + Mereo.* means that they listed both the constituents and the mereological object. In **P2**, the LLMs choose between the constituents (*Obj. 1* and *Obj. 2*), the mereological object (*Mereo. Obj.*) or any of the mentioned objects (*Any*). In **P3**, they answered Yes or No to whether the mereological object can be the referent. In **P4**, the LLMs all identified *it* as ambiguous. The *Ambiguous + Mereo.* column shows when the LLMs mention the mereological object as a possible referent.

**SAMPLE RESPONSE - AYA-EXPANSE-8B**

<USER> Consider the referent of the pronoun 'it' in the following sentence:  
 The hardworking, experienced autoworker removed the suspension from the chassis and sent it \_\_\_.  
 Can 'it' refer to the car?

<ASSISTANT> Yes, in the given sentence, the pronoun 'it' can refer to the car. The sentence is discussing an action performed on a car, specifically the removal of the suspension from the chassis. Since 'it' is used to refer to the object that is being acted upon, it logically follows that 'it' refers to the car.

Notably, these percentages are much higher than those in **P1** and **P2**. Yet, on the other hand, the difference between mereological and non-mereological contexts is quite small. Especially for Aya-Expanse-8B, the results are similar in the non-mereological cases, which is counter-intuitive because the mereological object should not be available as a referent given the detaching action. The text above is an example response where Aya-Expanse-8B thinks that *it* can refer to the mereological object in the non-mereological sentence. When the *car* (mereological object) is mentioned in the prompt, the model believes that the suspension and the chassis are parts of a car.

When the autoworker removes one of the parts (e.g., the suspension), the remaining is still a car. However, this line of reasoning does not appear in **P4**, when the mereological object is not mentioned in the non-mereological prompt. The model simply responds that both constituents can be the referent. This suggests a crucial difference in LLMs’ preference when a potential referent is mentioned in the prompt or not.

## 7 Discussion

In the above experiments, we assess LLMs’ ability to identify different types of coreference ambiguity using probability-based and prompting-based approaches. In the split-antecedent references, LLMs show a strong alignment with human preference (Moxey et al., 2004, 2012). When the roles assigned to the antecedents indicate that they may be involved in the same action, they are more likely to be referenced as a plural entity. However, in mereological contexts, the results are more complex. On the one hand, LLMs show preference for the mereological object following the verb, as shown in the *Object Production* experiment. On the other hand, the *Interpretation* experiment shows the lack



of self-consistency of LLMs between prompting strategies and hints about ambiguity and the mereological objects. Without hints of ambiguity, they almost never recognize the ambiguity when being prompted with **P1** and **P2**, although these prompts do not exclude the possibility that the pronoun is ambiguous. They also do not explicitly mention the mereological object as a referent except when it is provided in the prompt. Generally, LLMs prefer to link *it* to the first constituent.

However, in **P2** and **P3**, where the mereological objects are presented in the prompts, they are chosen or listed much more frequently (40%–60%). This leads to the possibility that LLMs may exhibit this psycholinguistic phenomenon when direct instruction is provided. This hypothesis is complemented by the results of the *Object Production* experiment where some LLMs (Aya-Expanse-8B and Qwen2.5-7B) give higher probabilities to mereological objects in mereological contexts. This possibility aligns with [Stengel-Eskin et al. \(2023\)](#) and [Karamolegkou et al. \(2025\)](#), who find that LLMs are better at producing different interpretations in few-shot contexts. Taken together, these results may imply that LLMs may not be able to apply their knowledge of ambiguity without specific instruction. In other words, they sometimes know the possible interpretations, but struggle to consistently show them in different types of instruction. This may be the case, as LLMs successfully detect ambiguity between the constituents in **P4**. However, the results of **P3**, where the difference between mereological and non-mereological contexts is not considerable, the mention of the mereological objects may not mean that the LLMs actually recognize the joined object. As mentioned before, the effect of frequency may be present. This suggests that the LLMs may be able to detect potential referents when they are mentioned in the input but fail to take into account the mereological object, a more complex referent. This suggests that while LLMs broadly capture the basic findings about plural pronoun preferential interpretation from classic papers such as ([Moxey et al., 2004, 2012](#)), their preferences differ from those captured in more recent literature such as [Cokal et al. \(2023\)](#).

We also observe that, while the LLMs perform quite similarly in the next-word prediction task, they exhibit different patterns in prompting tasks. Qwen2.5-7B seems to align more with humans. The model indeed chooses the mereological ob-

jects more in mereological contexts.

## 8 Conclusion

In this paper, we conduct a set of experiments to investigate whether LLMs are able to capture different interpretations of ambiguous plural reference and whether their representations of reference are comparable to how humans do in psycholinguistic literature. Our mixed findings contribute to a larger picture of the degree to which LLMs possess knowledge of ambiguity and how this knowledge is used.

## 9 Limitations

Our study provides new insights into how LLMs process text with different types of coreference ambiguity. Yet it has a few limitations that need to be noted. First, in the mereological cases, we think that the frequency of the constituents and the mereological object affect its likelihood to be chosen by the LLMs in both the next-token prediction and prompt tasks. For example, in the sentence *The shoemaker stripped the leather from the sole and inspected it \_\_*, the mereological object *shoe* is much more common than the *leather* or the *sole*. Although we try to select the verb that can be done on both constituents and the mereological objects, this effect of frequency is likely to be present, leading the LLMs to choose the mereological object as the referent without knowledge of the mereological relation. Secondly, the LLMs that we tested are rather small (around 7-8B). Investigations into larger LLMs are important as they may exhibit higher competency in interpreting ambiguous pronouns. In addition, we rely on prompting results, which may not always reflect the representation and reasoning capability of the LLMs. Future research may use recent mechanistic interpretability approaches for a more thorough and dimension-wise investigation of how LLMs represent ambiguous references ([Syed et al., 2024](#); [Vig et al., 2020](#); [Hanna and Mueller, 2025](#)). Such methods involve interventions on the LLMs' hidden activations, which potentially uncover the process underneath the LLMs' generation.

In addition, as pointed out by the reviewers, we may be quite strict in evaluating LLMs' ability to provide metalinguistic judgements about ambiguity because even humans can still have difficulty identifying it. We did not conduct the prompting experiments on humans, so we do not know their

preferences. We acknowledge this as one of the limitations of the paper.

## 10 Ethical Considerations

The study uses only publicly available linguistic stimuli from previous research and synthetic data generated by LLMs. It does not include any data from human participants.

## 11 Acknowledgments

We would like to thank three anonymous reviewers for their insightful comments. This project is funded by the Dutch Research Council (NWO) through the AiNed Fellowship Grant (Dealing with Meaning Variation, NGF.1607.22.002) to Massimo Poesio.

## References

- Jason E Albrecht and Charles Clifton. 1998. Accessing singular antecedents in conjoined phrases. *Memory & cognition*, 26(3):599–610.
- Nicholas Asher and Linton Wang. 2003. Ambiguity and anaphora with plurals in discourse. In *Semantics and Linguistic Theory*, pages 19–36.
- Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Derya Cokal, Ruth Filik, Patrick Sturt, and Massimo Poesio. 2023. Anaphoric reference to mereological entities. *Discourse Processes*, 60(3):202–223.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expand: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.
- Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The knowref coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961.
- Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665.
- Peter C Gordon, Randall Hendrick, Kerry Ledoux, and Chin Lung Yang. 1999. Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, 14(4):353–379.
- Michael Hanna and Aaron Mueller. 2025. Incremental sentence processing mechanisms in autoregressive transformer language models. 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 3181–3203.
- Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754.
- Antonia Karamolegkou, Oliver Eberle, Phillip Rust, Carina Kauf, and Anders Søgaard. 2025. Trick or neat: Adversarial ambiguity and language model evaluation. *arXiv preprint arXiv:2506.01205*.
- Sungryong Koh and Charles Clifton Jr. 2002. Resolution of the antecedent of a plural pronoun: Ontological categories and predicate symmetry. *Journal of Memory and Language*, 46(4):830–844.
- 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.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2023. We’re afraid language models aren’t modeling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807.
- Linda M Moxey, Anthony J Sanford, Patrick Sturt, and Lorna I Morrow. 2004. Constraints on the formation of plural reference objects: The influence of role, conjunction, and type of description. *Journal of Memory and Language*, 51(3):346–364.
- Linda M Moxey, Anthony J Sanford, and Karen Tonks. 2012. Representing characters in a scenario: What makes two individuals a set? *Language and cognitive processes*, 27(9):1405–1424.
- Nikole D Patson. 2014. The processing of plural expressions. *Language and Linguistics Compass*, 8(8):319–329.
- Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Massimo Poesio, Patrick Sturt, Ron Artstein, and Ruth Filik. 2006. Underspecification and anaphora: Theoretical issues and preliminary evidence. *Discourse processes*, 42(2):157–175.

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.

Anthony J Sanford and F Lockhart. 1990. Description types and method of conjoining as factors influencing plural anaphora: A continuation study of focus. *Journal of Semantics*, 7(4):365–378.

Anthony J Sanford and Linda M Moxey. 1995. Notes on plural reference and the scenario-mapping principle in comprehension. *Focus and coherence in discourse processing*, pages 18–34.

Anastasiia Sedova, Robert Litschko, Diego Frassinelli, Benjamin Roth, and Barbara Plank. 2024. To know or not to know? analyzing self-consistency of large language models under ambiguity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17203–17217.

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. 2023. Zero and few-shot semantic parsing with ambiguous inputs. *arXiv preprint arXiv:2306.00824*.

Aaquib Syed, Can Rager, and Arthur Conmy. 2024. Attribution patching outperforms automated circuit discovery. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 407–416.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401.

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.

Frank Wildenburg, Michael Hanna, and Sandro Pezzelle. 2024. Do pre-trained language models detect and understand semantic underspecification? ask the dust! In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9598–9613.

Juntao Yu, Nafise Sadat Moosavi, Silviu Paun, and Massimo Poesio. 2020. Free the plural: Unrestricted split-antecedent anaphora resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6113–6125.

## A Prompt Construction

### A.1 Prompt without Explicit Hints of Ambiguity

#### A.1.1 Prompt 1: Open-ended Questions

In the first prompt, we ask the LLMs to generate all possible antecedents for *it*. The goal is to see whether they have a deterministic choice or also consider other possible candidates.

#### Prompt template 1

```
<USER> Consider the pronoun "it" in the following sentence prefix. What is the referent of "it"? List all possible referents if there are any. Sentence: <sentence>

<ASSISTANT>
```

See the excerpt below for an example prompt and the LLM’s response.

```
<USER> Consider the referent of the pronoun 'it' in the following sentence: Sentence: The bike rider clipped the wheel to the frame and checked it __ What is the referent of 'it'? List all possible referents if any.
```

```
<ASSISTANT> In the sentence "The bike rider clipped the wheel to the frame and checked it," the pronoun "it" likely refers to either "the wheel" or "the frame."
```

Given the sequence of actions described, the most logical interpretation would be that the bike rider has just clipped the wheel to the frame and then proceeded to check the frame for proper attachment or alignment. However, without additional context, there’s also a possibility that "it" could refer to the wheel again, checking to ensure that the wheel has been securely attached to the frame.

Therefore, the possible referents for "it" are:

1. The wheel

## 2. The frame

We divide LLMs' responses into two categories. If the LLMs only list one of the constituents, their responses are coded as Ind.. If they choose between the two constituents, these are coded as Ind. + Ind. In case they mention both constituents and the mereological object, responses are coded as Ind. + Mereo.

### A.1.2 Prompt 2: Multiple-choice Questions

We append the expected pronoun to the sentence prefixes and ask the model to choose from the three objects using a multiple-choice format. To control for the effect of prompt difference, we use two types of prompts. In the multiple-choice question, there are two other options (*D. any of the above objects*), which imply that the pronoun is ambiguous and any of the mentioned objects can be the pronoun's referent and (*E. none of the above object*). Previous studies on how LLMs respond to multiple-choice questions have shown that LLMs prefer certain answer symbols, regardless of the actual content. In order to control for this effect, we shuffle the order of the options except for option D and E, resulting in a total of three different runs, and take the mean over runs.

#### Prompt template 2

```
<USER> Consider the pronoun "it" in the
following sentence:
Sentence: <sentence>
Which object does "it" refer to?
A. first object
B. second object
C. mereological object
D. any of the above objects
E. none of the above objects
Answer:
<ASSISTANT>
```

See the excerpt below for an example prompt and the LLM's response.

```
<USER> Consider the pronoun "it" in
the following sentence:
Sentence: John put the card into the en-
velope and hid it __.
Which object does "it" refer to?
A. the card
B. the envelope
C. the letter
```

- D. any of the above objects
- E. none of the above objects

Answer:

```
<ASSISTANT> The answer is A. the
card.
```

In the sentence "John put the card into the envelope and hid it," the pronoun 'it' refers to the last object mentioned before it, which is 'the card.' This is a common usage of pronouns in English, where 'it' is often used to refer back to the nearest noun in a sentence.

The answer choices are coded as follows.

- Obj. 1: The first object
- Obj. 2: The second object
- Mereo.: The mereological object
- Any.: Any of the above objects
- None.: None of the above objects

### A.2 Prompt with Explicit Mention of Ambiguity

#### A.2.1 Prompt 3: Open-ended Questions

We add various hints of ambiguity to study whether LLMs' responses are affected. To see if the LLMs are aware of the mereological object. We directly ask the LLMs if the mereological objects can be the referent for *it*. Responses are coded as either Yes or No.

#### Prompt template 3

```
<USER> Consider the pronoun "it" in the
following sentence. Can "it" refer to the
<mereological object>? Provide the answer
and explain.
Sentence: <sentence>

<ASSISTANT>
```

See the excerpt below for an example prompt and the LLM's response.

```
<USER> Consider the referent of the
pronoun 'it' in the following sentence:
Sentence: The shoemaker glued the
leather to the sole and inspected it __.
Can 'it' refer to the shoe?
```

```
<ASSISTANT> Yes, "it" can refer to the
shoe in this sentence. The context sug-
gests that after the shoemaker has glued
```



the leather to the sole, they then inspect the completed shoe or the work as a whole. Without additional context, the most natural and probable interpretation is that "it" refers to the shoe that has been recently made. However, depending on the broader context of the paragraph or the surrounding sentences, there could potentially be other referents for "it."

### A.2.2 Prompt 4: Open-ended Questions

Finally, we use an open question format to test whether LLMs capture the fact that the pronoun *it* can refer to multiple referents. As shown in prompt 2, the model was asked to tell whether the pronoun *it* is ambiguous and explain their answer. <sentence> is the sentence prefix.

#### Prompt template 4

```
<USER> Is the pronoun "it" in the following sentence ambiguous? Provide the answer and explain.  
Sentence: <sentence>  
  
<ASSISTANT>
```

See the excerpt below for an example prompt and the LLM's response.

```
<USER> Is the pronoun 'it' in the following sentence ambiguous?  
Provide the answer and explain.  
Sentence: The railway worker attached the engine to the goods wagon and sent it __.  
  
<ASSISTANT> Yes, the pronoun 'it' in the given sentence is ambiguous.
```

The sentence "The railway worker attached the engine to the goods wagon and sent it \_\_" does not clearly specify whether 'it' refers to the 'engine' or the 'goods wagon'. Without additional context, there are two possible interpretations:

1. The railway worker attached the engine to the goods wagon and then sent (moved) the engine.
2. The railway worker attached the engine to the goods wagon and then sent (moved) the goods wagon.

The ambiguity arises because 'it' could be referring to either of the two nouns mentioned previously in the sentence without any clear indicators such as an article or modifying phrase. To remove this ambiguity, more context or specific language would be needed to clarify what exactly was sent.

LLMs responses are divided into: (1) Unambiguous, where the LLMs do not recognize the ambiguity of the pronoun, (2) Ambiguous, where the LLMs think that the two constituents are both possible to the referent, and (3) Ambiguous + Mereo., where they also mention that the mereological object is also likely to be a referent.