Solving the Challenge Set without Solving the Task: On Winograd Schemas as a Test of Pronominal Coreference Resolution

Ian Porada
Mila - Quebec AI Institute
McGill University
ian.porada@mail.mcgill.ca

Jackie Chi Kit Cheung
Mila - Quebec AI Institute
McGill University
Canada CIFAR AI Chair
jackie.cheung@mcgill.ca

Abstract

Challenge sets such as the Winograd Schema Challenge (WSC) are used to benchmark systems' ability to resolve ambiguities in natural language. If one assumes as in existing work that solving a given challenge set is at least as difficult as solving some more general task, then high performance on the challenge set should indicate high performance on the general task overall. However, we show empirically that this assumption of difficulty does not always hold. In particular, we demonstrate that despite the strong performance of prompted language models (LMs) on the WSC and its variants, these same modeling techniques perform relatively poorly at resolving certain pronominal ambiguities attested in OntoNotes and related datasets that are perceived to be easier. Motivated by these findings, we propose a method for ensembling a prompted LM with a supervised, task-specific system that is overall more accurate at resolving pronominal coreference across datasets. Finally, we emphasize that datasets involving the same linguistic phenomenon draw on distinct, but overlapping, capabilities, and evaluating on any one dataset alone does not provide a complete picture of a system's overall capability.

1 Introduction

The Winograd Schema Challenge (WSC; Levesque et al., 2012) is a challenge set of ambiguous pronominal coreference resolution (PCR) problems, one of many popular challenge sets used to evaluate NLP systems (*e.g.,* Isabelle et al., 2017; Clark et al., 2018; McCoy et al., 2019). Challenge sets are constructed to consist of relatively difficult instances of some more general task. In many cases, systems' performance on challenge sets is considered in isolation of performance on the broad range of ambiguous expressions attested in natural corpora on which the general task being studied could also be evaluated;[1] *e.g.,* systems are often evaluated on the WSC without considering how those same systems might perform on a diverse range of attested pronominal expressions (Kocijan et al., 2019b; Shen et al., 2021; Gao et al., 2023; Achiam et al., 2023, *i.a.*).

Constructed WSC Pair

Jim yelled at Kevin because he was so upset.
Jim comforted Kevin because he was so upset.

Attested Pronominal Expression

... Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came down on Monday as so much delighted with it, that he agreed with Mr. Morris immediately; ...

Figure 1: *Top*: An example minimal pair from the WSC. *Bottom*: Pronouns attested in the novel *Pride and Prejudice* and annotated for coreference by Vala et al. (2016).

The WSC specifically consists of minimal pairs of sentences, each containing an ambiguous pronoun (Figure 1). These pairs are manually constructed such that consistently disambiguating the pronouns is believed to require the types of commonsense world knowledge and reasoning ability a human reader might rely on. Considering the recent success of language model (LM) based approaches at resolving WSC instances, some of the original authors of the WSC have declared the challenge set solved (Kocijan et al., 2023).

And yet, in this work we demonstrate that the same LM-based systems that have reportedly solved the WSC and its variants are relatively inaccurate at resolving certain ambiguous pronominal expressions attested in natural corpora and annotated in OntoNotes (Hovy et al., 2006) and related

[1] We use the term *natural corpora* to refer to text that was not explicitly constructed or elicited for research purposes. An *attested* expression is one appearing in natural corpora in contrast to *constructed* expressions that commonly compose challenge sets.

Proceedings of the 28th Conference on Computational Natural Language Learning, pages 489–506
November 15-16, 2024 ©2024 Association for Computational Linguistics

datasets. One may find this result surprising given that "the point of the WSC is to test programs that claim to have solved the problem of pronoun reference resolution" (Kocijan et al., 2023) and that WSC instances are believed to represent relatively difficult examples of PCR (Peng et al., 2015).

We specifically consider prompted LMs as systems that are relatively accurate at resolving Winograd schemas; among LMs, our experiments focus on the Llama family of models (Touvron et al., 2023; Dubey et al., 2024), although we present evidence that our results generalize across LM families including OLMo (Groeneveld et al., 2024) and Mistral (Jiang et al., 2023). We compare the performance of prompted LMs up to 70B parameters against state-of-the-art coreference resolution systems, such as Maverick (Martinelli et al., 2024), which are known to be accurate at resolving attested pronominal coreferences.

We evaluate systems across 11 datasets. Six of these datasets contain PCR problems for text attested in natural corpora, *e.g.,* OntoNotes 5.0 (Weischedel et al., 2013) and OntoGUM (Zhu et al., 2021). The other five datasets consist of PCR problems that were constructed for WSC-like challenge sets, *e.g.,* Winogrande (Sakaguchi et al., 2021) and DPR (Rahman and Ng, 2012).

When comparing against unsupervised baselines, we find LMs are generally more accurate across all datasets; however, *whereas supervised coreference resolution models perform relatively poorly on the WSC, we find these same systems are more accurate than prompted LMs at resolving certain attested pronouns*. This finding is consistent across test sets of diverse annotation guidelines and textual genres.

Motivated by these results, we propose a method for ensembling a prompted LM with a task-specific system in order to achieve a final system that is overall more accurate at resolving pronominal coreference across datasets. This ensembling method functions by heuristically determining salient discourse entities for which coreference is disambiguated by a state-of-the-art coreference resolution system trained on OntoNotes. Meanwhile, the remaining instances are disambiguated using an LM prompted with in-context examples. In most cases, the final system is more accurate at resolving pronouns occurring in attested expressions and WSC-like challenge sets.

Ultimately, our findings illustrate the point that datasets involving the same linguistic phenomenon draw on distinct, but overlapping, capabilities; therefore, no one dataset alone is capable of providing a complete picture of a system's overall performance. We therefore argue that challenge set results should be considered in conjunction with results on evaluations that encompass a diverse range of attested expressions.

Contributions. Our primary contributions can be summarized as follows:

1. We formalize and empirically question *the challenge set assumption* that solutions to a challenge set generalize to diverse, attested instances of the phenomenon being targeted. In the case of PCR, we provide direct evidence that this assumption does not hold.
2. We present a formatted collection of 11 datasets that follow the same, consistent formulation of PCR. Using this collection, we evaluate and compare multiple types of approaches to PCR including supervised models, prompted LLMs, and rule-based systems.

2 Related Work

PCR is broadly the task of determining which linguistic expressions refer to the same discourse entity as a given pronominal expression (Hobbs, 1978). See Zhang et al. (2021) and Poesio et al. (2023) for related surveys.

Proposed systems for resolving pronominal coreference have traditionally relied on heuristic rules often in combination with unsupervised statistical patterns of handcrafted features (Poon and Domingos, 2008; Charniak and Elsner, 2009; Raghunathan et al., 2010; Lee et al., 2011, *i.a.*). More recently, LM-based approaches have been proposed including: LMs finetuned on supervised training data (Zhang et al., 2019c; Zhao et al., 2022), weakly supervised LMs (Kocijan et al., 2019a; Shen et al., 2021), and prompting LMs by formatting PCR as either a cloze task (Trinh and Le, 2018; Radford et al., 2019) or question answering (Brown et al., 2020; Wang et al., 2022; Le and Ritter, 2023; Zhu et al., 2024).

The ability of a system to perform PCR has been evaluated generally on: 1) collections of ambiguous pronouns attested in natural text (Hobbs, 1978; Lappin and Leass, 1994; Webster et al., 2018), 2) the subsets of larger coreference resolution datasets that include pronominal coreference (Martschat and Strube, 2014; Zhang et al., 2019c; Lu and Ng, 2020), and 3) challenge sets composed of WSC-

like instances (Rahman and Ng, 2012; Emami et al., 2019; Sakaguchi et al., 2021).

The WSC and inspired datasets have been adopted by researchers studying the more general task of coreference resolution to be used as challenge sets in addition to more canonical evaluations such as OntoNotes (Peng et al., 2015; Toshniwal et al., 2021; Zhao et al., 2022). Such work has shown that systems designed for coreference resolution perform poorly on the WSC. We adopt the perspective of this line of work and view WSC-like datasets as challenge sets of PCR.

Recent advances in language modeling have proven accurate at resolving WSC instances when compared to earlier approaches, in some cases nearing approximates of human accuracy (Brown et al., 2020; Wei et al., 2022; Touvron et al., 2023). However, similar techniques have been shown to be less accurate than supervised models when evaluated on established evaluations of the general task of coreference resolution (Yang et al., 2022; Le and Ritter, 2023; Zhu et al., 2024; Gan et al., 2024). Our work diverges from these studies by focusing specifically on PCR rather than the broader concept of coreference which has multiple competing definitions (Recasens and Hovy, 2010; Zeldes, 2022).

3 Method

In this section, we formulate the problem of pronominal coreference resolution (PCR) and provide a high-level description of how system accuracy is evaluated. We also formalize the assumptions commonly made when evaluating on challenge sets so that we can explicitly test if these assumptions hold.

3.1 Problem Formulation

We consider the task of PCR formulated as follows: given a text passage $w = (w_1, \ldots, w_t)$, resolve some pronominal expression x to its correct antecedent a, where x and a are subspans of w. We study a restricted version of this problem formulated as binary classification.

More explicitly, we assume that exactly one of two candidate antecedents in w is the correct resolution of x. This formulation accommodates both WSC-style and datasets containing annotations of coreference in occurring in natural corpora. Formally, given w, x, and a set of two candidate antecedents $\{\hat{a}_1, \hat{a}_2\}$, the goal is to correctly determine which candidate antecedent corresponds to the true antecedent $a = w_{k:l}$. The other candidate is some distractor mention $b = w_{m:n}$ that refers to a discourse entity but does not corefer with x. An example instance is given in Figure 2.

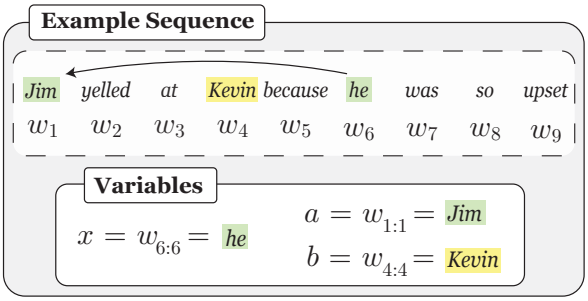

Figure 2: An example instance and the corresponding variables: the pronoun x, antecedent a, and distractor candidate b.

3.2 Challenge Set Assumptions

An assumption of the WSC is that solving WSC instances is more difficult than resolving other instances of the PCR task such as pronouns attested in natural corpora. We formulate this assumption as follows (Def. 1). We will then test this assumption empirically by comparing the performance of various systems across both challenge set instances and attested pronouns.

To premise, let C be a challenge set and D some other dataset representing the same task. Furthermore, let θ and ϕ be systems that are to be evaluated based on their performance on the given task. Function U represents a measure of the performance of a system on a given dataset, *e.g.,* the performance of θ on C is measured as $U(\theta, C)$.

Definition 1 (The Challenge Set Assumption) *The ordering of model performance on the challenge set C is preserved on dataset D. That is,* $U(\theta, C) > U(\phi, C) \implies U(\theta, D) > U(\phi, D)$.

Intuitively, the assumption is that because C is strictly more difficult than D, systems that are relatively accurate on C should be relatively accurate on D as well.

3.3 Evaluating Performance

To test this assumption, we evaluate systems across multiple test sets. Here we describe how the performance function U is calculated.

Attested Pronominal Expressions. To evaluate performance on attested pronominal expressions, we start with some existing dataset of identity coreference relations annotated in datasets of curated

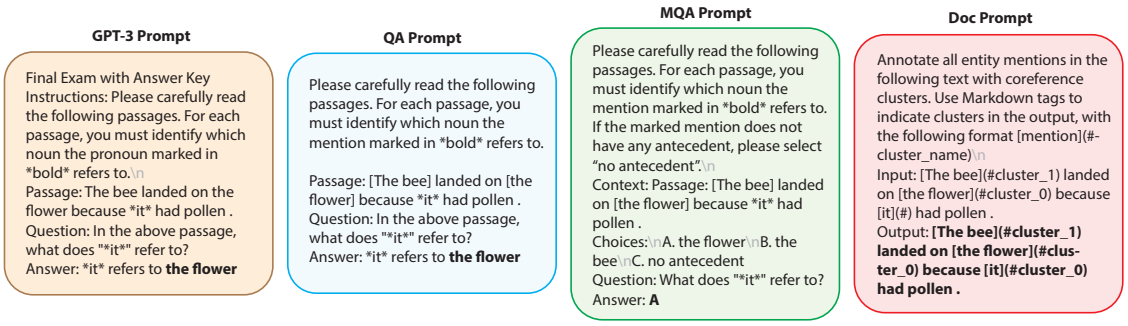

Figure 3: A training set instance from the Definite Pronoun Resolution (DPR) dataset (Rahman and Ng, 2012) formatted using each of the corresponding prompts. Denoted in bold is the expected model output. The GPT-3 prompt (Brown et al., 2020) does not rely on gold mention span annotations. QA Prompt and Doc Prompt were presented by Le and Ritter (2023). The multiple-choice QA (MQA) prompt was presented by Zhu et al. (2024).

natural corpora. We then take all mentions that are in a coreference relation and are within a predefined set of pronouns to be a coreferring pronominal expressions x.[2] We take the text passage w to be the concatenation of the sentence in which x occurs with the preceding two sentences. For each x for which a single coreferring nominal antecedent mention a occurs in the context w, and for which at least one coreferring expression b that does not corefer with x also occurs in w, we create a test instance. In the event that there are multiple candidates that could be chosen as b, we randomly sample one. We measure performance based on a system's accuracy at resolving these instances.

This formulation and the predefined set of pronouns follow the conventional setup for PCR used in existing work (Yang et al., 2003; Ng, 2005; Li et al., 2011; Zhang et al., 2019c; Zhao et al., 2022). App. C provides further details regarding how datasets are formatted.

WSC-like Challenge Sets. WSC instances generally follow the formulation of PCR we have outlined above—*i.e.,* the basic premise of a WSC is that there is some text w containing a pronoun x with two candidate antecedents $\{\hat{a}_1, \hat{a}_2\}$—so we perform minimal formatting of existing WSC-like challenge sets so that examples are in the same form as for attested datasets. This requires tokenization and in certain cases determining the exact span of candidate mentions. Further details are provided in App. C. We can then directly compute accuracy as the ratio of instances where the system predicts the correct candidate antecedent.

[2]The following strings are considered as pronouns: "she", "her", "he", "him", "them", "they", "it", "his", "hers", "its", "their", "theirs", "this", "that", "these", "those".

4 Experiments

In this section, we describe our experimental setup in detail which tests whether the challenge set assumption holds empirically. We compare prompted LMs that are known to be accurate at the WSC against task-specific systems known to be accurate at resolving certain attested pronouns (§4.1). Systems are evaluated across 11 datasets spanning both attested and WSC-like instances (§4.2).

4.1 Systems

4.1.1 Prompted Language Models

In recent years, prompted LMs have proven accurate at the WSC. One would therefore expect such systems to be relatively accurate at resolving attested pronominal expressions if the challenge set assumption holds. Prompted LMs function by predicting the correct antecedent span a given a problem instance $(w, x, \{\hat{a}_1, \hat{a}_2\})$ which is formatted using a particular textual prompt template that may possibly include in-context examples.

Llama 3.1 As a prompted LM we focus on the Llama 3.1 family of models at various sizes (Dubey et al., 2024). These are competitive open-weights LMs. We consider either the base or instruct version as specified in each experiment. The instruct versions were additionally finetuned on instruction-tuning data, such as the Flan collection (Longpre et al., 2023), and human preference annotations. We evaluate the 8B and 70B parameter model sizes.

In the experiments where we consider few-shot prompted Llama models, we also compare against a supervised Llama 3.1 8B model which we finetune to resolve WSC by training on public training sets formatted using a QA prompt.

Additional LMs We additionally compare performance against the smaller Llama 3.2 models, the fully open source OLMo model (Groeneveld et al., 2024), and the Mistral-NeMo 12B parameter model (Mistral AI Team, 2024).

Prompting Techniques. Our goal is not to propose new prompting techniques, so we experiment using four existing prompt templates sourced from the literature. These templates are shown in Figure 3. The GPT-3 prompt was used by Brown et al. (2020) for evaluating GPT-3 on the SuperGLUE WSC (Wang et al., 2019) and does not require gold mention annotations. For this prompt, we check the string match of the model output as in Brown et al. (2020). The additional prompts (QA, MQA, and Doc prompts) were proposed for using language models to explicitly perform the task of coreference resolution and do require explicit candidate mention spans. For these prompts, of the candidate outputs, we take that with the highest probability assigned by the language model to be the model prediction. Another common approach is to formulate WSC-like instances as a cloze-task (Trinh and Le, 2018; Gao et al., 2023). We do not consider this prompting technique, however, as it is not compatible with pronominal references whose resolution depends on the grammatical features of the pronoun being considered.

We evaluate prompted LMs in zero- and few-shot settings depending on what comparison is being made. In the zero-shot setting, the LM is only prompted with the corresponding instructions and input passage. In the few-shot setting, we use instruction-tuned version of the Llama 3.1 models with 32 training instances prepended to the input.

4.1.2 Task-Specific Systems

We compare the performance of prompted LMs against the following task-specific systems designed for the general problem of coreference resolution. Such coreference resolution models are believed to perform poorly on the WSC. One would therefore expect prompted LMs to outperform these task-specific systems across all PCR datasets given *the challenge set assumption* (Def. 1).

dcoref As a representative unsupervised system, we consider the "Stanford Deterministic Coreference Resolution System" (dcoref; Lee et al., 2013). This is a deterministic, rule-based approach to the general problem of identity coreference resolution and does not rely on supervised examples of coreference relations. The system is optimized to perform well on the OntoNotes dataset. This system uses 10 sieves (such as string match and grammatical feature agreement) to identify potentially coreferring mentions. We use the most recent version implemented in Stanford Core NLP (Manning et al., 2014). Around 30 percent of OntoNotes errors are described as pronominal anaphora errors in the original dcoref paper.

Maverick As a representative example of a supervised system, we consider the state-of-the-art Maverick coreference resolution system (Martinelli et al., 2024). We use the publicly released weights of the best performing system which consists of a DeBERTa-v3 encoder (He et al., 2021) finetuned on OntoNotes.

4.2 Datasets

We evaluate systems on 11 datasets including curated datasets of pronouns attested in natural corpora, such as OntoNotes (Weischedel et al., 2013), and challenge sets of WSC-like instances, such as the original WSC test set (Levesque et al., 2012) and DPR (Rahman and Ng, 2012).

4.2.1 Attested Pronominal Expressions

As noted in the methods section (§3), our tests that involve attested pronouns are based on restricting the set of annotations in more general coreference resolution datasets to be evaluated as binary classification problems similar to the WSC. The datasets that we use to achieve this are described below. Four of these are datasets of nominal identity coreference annotated in English-language, document-level passages (OntoNotes, OntoGUM, PreCo, and ARRAU). The remaining two focus exclusively on PCR (GAP and PDP).

OntoNotes OntoNotes 5.0 (Weischedel et al., 2013) consists of seven genres including news, conversations, and web data annotated for coreference by two experts. This dataset has been used in prior work to explicitly evaluate PCR both in isolation (Zhang et al., 2019c, 2021, 2019b) as well as PCR as a failure case of more general coreference resolution systems (Lu and Ng, 2020). We use the standard English CoNLL-2012 Shared Task version of this dataset (Pradhan et al., 2012).

OntoGUM OntoGUM (Zhu et al., 2021) is a reformatted version of the GUM corpus (Zeldes, 2017) which was annotated for coreference by

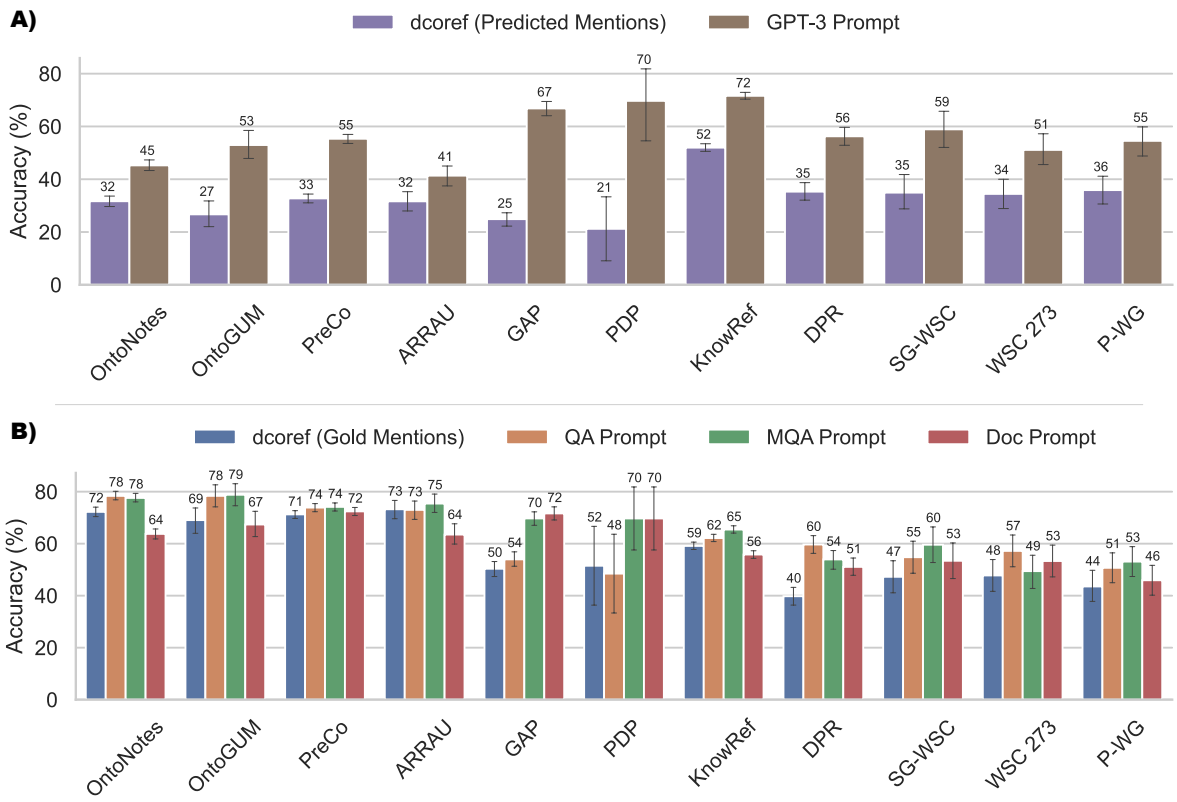

Figure 4: A comparison of the rule-based dcoref system (Lee et al., 2013) and the Llama 3.1 8B base model prompted for PCR using various prompts. **A)** Systems that do not need gold mention spans. Across datasets, Llama 3.1 with the GPT-3 prompt always outperforms the dcoref baseline. **B)** Systems that require gold mention spans as input. In general, prompted Llama 3.1 is more accurate than dcoref on both attested and constructed instances.

linguistic students. We use version 9.2.0 of OntoGUM. This dataset is designed to follow the same annotation guidelines as OntoNotes while expanding coverage to additional textual genres such as web forums and video blogs.

PreCo PreCo (Chen et al., 2018) is a large-scale dataset of English exams annotated for coreference.

ARRAU ARRAU 2.1 (Uryupina et al., 2020) is a dataset of written news and spoken conversations annotated for various anaphoric phenomena by experts. We use the version formatted by Xia and Van Durme (2021). The annotation guidelines differ from OntoNotes, and additional phenomenon have been annotated including extensive semantic and syntactic features of mentions.

GAP GAP (Webster et al., 2018) is a dataset of pronouns attested in English Wikipedia and annotated for coreference. We study only instances where exactly one of two candidate antecedents is coreferring with the given pronoun to match our PCR problem formulation.

PDP PDP (Morgenstern et al., 2016) is a collection of 80 pronoun disambiguation problems attested in text which was used for the original version of the WSC in order to test systems on examples believed to be relatively easy.

4.2.2 WSC-like Challenge Sets

The five challenge sets that we evaluate on are as follows. To standardize the format of these datasets, we consider the lexical units w_i to be syntactic words. We split the raw text into these syntactic words using the Stanza library (Qi et al., 2020).

KnowRef-60K Emami et al. (2020) presented WSC-like instances which were created by perturbing internet forum text using heuristic rules. Thus, these instances fall somewhere in between attested and constructed.

DPR Definite pronoun resolution (DPR; Rahman and Ng, 2012) is a dataset of instances in a similar format to the original WSC without the strict requirement that instances cannot be resolved based on simple selectional preferences.

SuperGLUE WSC (SG-WSC) The set of WSC instances used for the SuperGLUE benchmark (Wang et al., 2019) which was originally modified from WSC 273 and PDP.

WSC 273 The original WSC (Levesque et al., 2012) consisting of 273 instances. We manually annotated mentions to fit our format similar to as in McCann et al. (2018) and Toshniwal et al. (2021).

Pronominal Winogrande (P-WG) We use the portion of the Winogrande test set (Sakaguchi et al., 2021) which contains person entities. We replace the underscore with an appropriate third-person pronoun as in Porada et al. (2023).

5 Results

We first present results comparing zero-shot prompting methods with the unsupervised dcoref system. We then compare the best performing prompting method against the supervised Maverick coreference resolution system and a supervised Llama 3.1 baseline. Across all figures, error bars represent 90% confidence intervals. Results are presented on the corresponding test splits using the best model configuration. Additional details are presented in App. D.

Comparing prompted LMs against earlier unsupervised systems, the challenge set assumption does hold. Results for the fully unsupervised systems are presented in Figure 4. We observe that generally prompted LMs, which outperform dcoref on the WSC variants, also outperform dcoref on datasets of attested pronominal expressions. We also see that model performance is sensitive to the prompt format.

Exceptions are on the PDP dataset, whose small size makes it difficult to draw generalizable conclusions, and in the case of the Doc Prompt, which has high variance across datasets. Le and Ritter (2023) similarly found that Llama models did not consistently generalize with the Doc Prompt template.

In Figure 5, we compare accuracies of various LMs using the QA prompt template. Our conclusion, that prompted LMs outperform dcoref on both constructed and attested instances, is consistent across LM families.

However, when comparing prompted LMs against a supervised coreference resolution system, the challenge set assumption does not hold. In Figure 6, we present the accuracy of Llama 3.1 70B using the QA prompt in a few-shot setting compared against a supervised coreference resolution system. While the prompted LM is more accurate across WSC-like datasets (with the exception of KnowRef), the supervised coreference resolution system is more accurate at resolving attested pronominal coreferences.

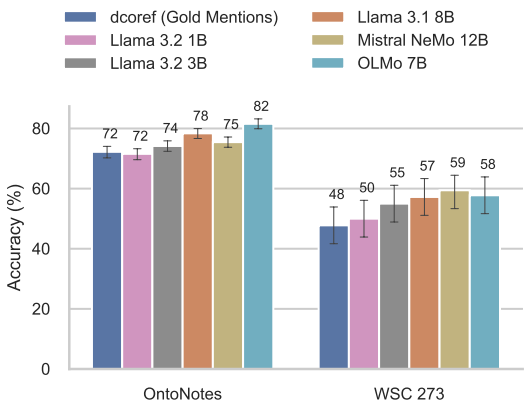

Figure 5: Accuracies of various LMs using the QA prompt template as compared against a dcoref baseline. We find that LMs generally outperform dcoref on both attested and constructed instances.

For this experiment, we consider the instruction-tuned version of Llama 3.1 with 32 in-context examples as a prototypical example of an LM as evaluated on WSC-like datasets. When we compare against a supervised Llama 3.1 base model, trained on the Winogrande and DPR training sets for 5k steps, the difference in accuracies across datasets is even more extreme.

The exceptional case of KnowRef may be due to the fact that this dataset is constructed by perturbing attested pronominal expressions and may be overall more similar to collections of attested rather than constructed linguistic expressions.

6 Analysis

Our results thusfar do not answer the question of *why* the challenge set assumption does not hold. Heuristic estimates of features such as number and animacy are typically required to agree between an antecedent and a pronoun in order for the two to be coreferring, but these features are never required to resolve WSC instances per their design. Therefore, it may be the case that prompted LMs are not sufficiently considering these features for the attested PCR problems. To test this hypothesis, we analyze to what extent LMs could benefit from, or already incorporate implicitly, the use of such

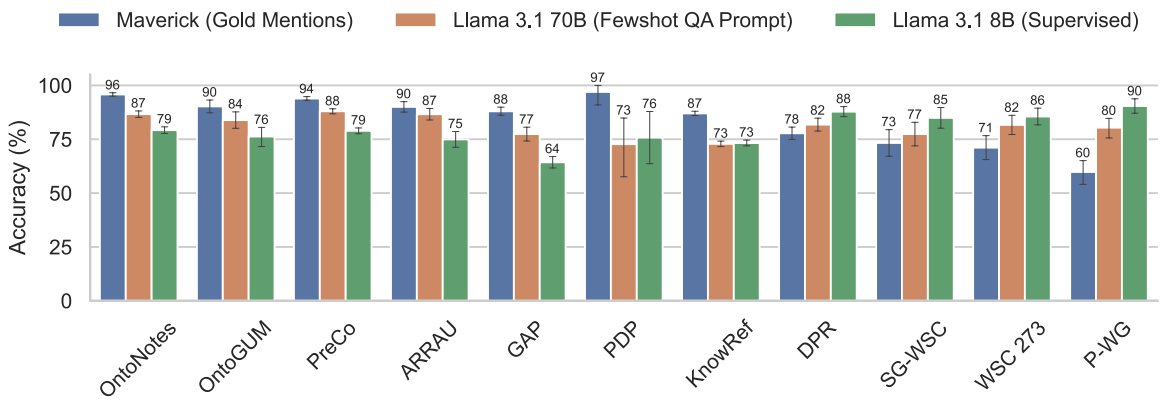

Figure 6: A comparison of the accuracy of Llama 3.1 70B instruct (32-shot) against the supervised Maverick coreference resolution system. We observe that the challenge set assumption does not hold; that is, despite being generally more accurate on WSC-like datasets, the prompted LM is less accurate on datasets of pronominal expressions attested in natural corpora. From left to right, the first six datasets consist of attested examples, and the remaining five are WSC-like challenge sets.

features. To do so, we experiment with oracle baselines including these features in the model input as a verbalized statement.

6.1 Verbalized Features

The verbalized features that we consider are those annotated in the ARRAU corpus. These are: 1) grammatical gender, 2) number, 3) enamex type (i.e., semantic type: is the entity a person, organization, or location?), and 3) distance between mentions. We also explore incorporating a gold, annotated label as an oracle baseline.

Prompt. Verbalized features are appended to the input string in the form:

> The [FEATURE_NAME] of "[X]" is [Y].

For example, the passage in Figure 2 is prepended with verbalizations such as:

> The grammatical gender of "Jim" is male.

Results of this experiment are presented in Table 1. We observe some accuracy increase from the inclusion of grammatical gender, but otherwise no influence. Meanwhile, the oracle baseline suggests models are capable of incorporating verbalized features that perfectly align with the correct antecedent prediction (i.e., gold labels).

	ARRAU
Llama 3.1 70B (QA Prompt)	0.86
+ gold gender	0.87
+ gold number	0.85
+ gold enamex type	0.86
+ distance between mentions	0.84
+ gold label (oracle)	0.99

Table 1: Results including additional features in the model input on the ARRAU validation set.

7 Ensembling Systems for Better Performance

Finally, we present results for our proposed ensembling method. This method is motivated as follows: because the challenge set assumption does not hold, prompted LMs and task-specific systems have distinct strengths at PCR.

7.1 Method

This method functions by heuristically determining if x corresponds to a salient discourse entity, in which case a supervised coreference resolution system is used to predict the correct antecedent a. Otherwise, x is resolved using a prompted LM. This approach benefits from the fact that supervised coreference resolution systems are relatively accurate at resolving pronominal expressions that corefer to the most salient discourse entities. Meanwhile, prompted LMs are relatively accurate at resolving pronominal expressions referring to infrequently mentioned entities.

7.2 Implementation

For our proposed ensembling method, we first predict pronominal coreferences using both the supervised Maverick system and the prompted Llama 3.1 70B instruct model as before. We then heuristically

System	OntoGUM	PreCo	WSC 273
Maverick	0.90	0.94	0.71
Llama 3.1 70B	0.84	0.88	0.82
Ensemble	0.90	0.95	0.84

Table 2: Accuracy of the ensemble method compared against Maverick (supervised coreference resolution) and prompted Llama 3.1 70B instruct.

determine if a candidate antecedent corresponds to a salient discourse entity based on the number of coreferring noun phrases predicted by Maverick in an end-to-end setup. When the number of predicted coreferring mentions is greater than two (that is, the pronoun is estimated to corefer to more than one other linguistic expression) we use the Maverick predictions given gold mention spans. Otherwise, we use the Llama predictions.

7.3 Results

We present the results for our ensembling method on three out-of-domain datasets of attested pronominal expressions in Table 2. The ensemble predictions are at least as accurate as the best performing model, and in the case of PreCo and WSC 273, more accurate than the single most accurate system. Results across all datasets are presented in App. E.

8 Discussion

Coreference resolution systems have traditionally struggled at resolving pronouns when the resolution depends on semantic knowledge related to high-order predicate-argument relations (Kehler et al., 2004; Durrett and Klein, 2013; Zhang et al., 2019a). Meanwhile, our results suggest that resolving WSC instances, which are designed to explicitly rely on such knowledge, is in some ways relatively easier than other cases for prompted LMs. Therefore, our intuitions as a research community regarding what constitutes challenging examples may not always be aligned with the actual failure cases of newer modeling paradigms. Consequently, we must be careful as a community to not interpret high performance on challenge sets as indicating that the more general task being studied can consistently be solved by a given system.

The Solvability of PCR. Our experiments and results are not intended to make claims regarding the solvability of the task of PCR. It may be that alternative prompting formats exist for which Llama models are relatively more accurate at resolving attested pronominal coreferences, and one would expect accuracy to increase with LM size. What our results do show, rather, is that existing approaches that are successful on the WSC and variants cannot generalize to all attested PCR problems.

Coreference and Substitutability. By their design, WSC instances can be formatted as a cloze-style task where the correct antecedent is that which is most likely to be substituted for the ambiguous pronoun. Substitutability and coreference are related but distinct concepts, however. While WSC instances are difficult in that they cannot be solved with the agreement of features between a pronoun and a candidate antecedent, they differ from some attested PCR problems in that for WSC instances the concept of coreference is aligned with substitutability. One possible hypothesis to explain our results is that this alignment is useful for solving the WSC. This hypothesis is based on the idea that substitutability can be formatted as a cloze-style task and is therefore closely aligned with the LM pretraining objective.

Data Contamination. An open question is whether LMs are exposed to the WSC or other datasets' test instances during pretraining. Elazar et al. (2023) estimate that up to 30% of WSC test instances may be contaminated in the training corpus of Llama and other language models. However, OntoNotes, OntoGUM, Winogrande, Knowref, and GAP are estimated to have close to zero contamination according to the Data Contamination Database (CONDA Workshop Organizers, 2024). ARRAU is not publicly distributed and also unlikely to be contaminated. Because our results are consistent for datasets that are likely not contaminated, we believe that issues of data contamination are unlikely to invalidate our findings.

9 Conclusion

The ability to disambiguate pronominal expressions is necessary for interpreting natural language and has been used extensively as a benchmark to evaluate models of semantics and discourse.

In this work, we study several possible approaches to modeling pronominal coreference. Across evaluations, we find that prompting a large language model (LM) outperforms other approaches on the WSC, but underperforms on certain attested occurrences of pronouns annotated for coreference in OntoNotes and related datasets.

Acknowledgements

Ian Porada is supported by a doctoral fellowship from the Fonds de Recherche du Québec Nature et Technologies (FRQ-NT). This research was enabled in part by compute resources provided by Mila (mila.quebec). We thank the anonymous reviewers for their valuable suggestions which greatly improved this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 148–156, Athens, Greece. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

CONDA Workshop Organizers. 2024. Data contamination database.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, et al. 2024. The llama 3 herd of models.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA. Association for Computational Linguistics.

Yanai Elazar, Akshita Bhagia, Ian Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, et al. 2023. What's in my big data? *arXiv preprint arXiv:2310.20707*.

Ali Emami, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2020. An analysis of dataset overlap on Winograd-style tasks. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5855–5865, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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, Torino, Italia. ELRA and ICCL.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Tamanna Hossain, Sunipa Dev, and Sameer Singh. 2023. MISGENDERED: Limits of large language models in understanding pronouns. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5352–5367, Toronto, Canada. Association for Computational Linguistics.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Pierre Isabelle, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496, Copenhagen, Denmark. Association for Computational Linguistics.

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

Andrew Kehler, Douglas Appelt, Lara Taylor, and Aleksandr Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 289–296, Boston, Massachusetts, USA. Association for Computational Linguistics.

Vid Kocijan, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019a. WikiCREM: A large unsupervised corpus for coreference resolution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4303–4312, Hong Kong, China. Association for Computational Linguistics.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019b. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The defeat of the winograd schema challenge. *Artificial Intelligence*, 325:103971.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers?

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA. Association for Computational Linguistics.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*.

Dingcheng Li, Tim Miller, and William Schuler. 2011. A pronoun anaphora resolution system based on factorial hidden Markov models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1169–1178, Portland, Oregon, USA. Association for Computational Linguistics.

Tyne Liang and Dian-Song Wu. 2003. Automatic pronominal anaphora resolution in English texts. In *Proceedings of Research on Computational Linguistics Conference XV*, pages 111–127, Hsinchu, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Jing Lu and Vincent Ng. 2020. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online. Association for Computational Linguistics.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. 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.

Sebastian Martschat and Michael Strube. 2014. Recall error analysis for coreference resolution. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2070–2081, Doha, Qatar. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Mistral AI Team. 2024. Mistral nemo.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy. Association for Computational Linguistics.

Leora Morgenstern, Ernest Davis, and Charles L Ortiz. 2016. Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, 37(1):50–54.

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *AAAI*, pages 1081–1086.

Haoruo Peng, Daniel Khashabi, and Dan Roth. 2015. Solving hard coreference problems. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 809–819, Denver, Colorado. Association for Computational Linguistics.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9:561–587.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 650–659, Honolulu, Hawaii. Association for Computational Linguistics.

Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. 2023. Investigating failures to generalize for coreference resolution models. *arXiv preprint arXiv:2303.09092*.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: The Winograd schema challenge. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 777–789, Jeju Island, Korea. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2010. Coreference resolution across corpora: Languages, coding schemes, and preprocessing information. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1423–1432, Uppsala, Sweden. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Ming Shen, Pratyay Banerjee, and Chitta Baral. 2021. Unsupervised pronoun resolution via masked noun-phrase prediction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 932–941, Online. Association for Computational Linguistics.

Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. 2021. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Trieu H. Trinh and Quoc V. Le. 2018. A simple method for commonsense reasoning.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J. Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Hardik Vala, Stefan Dimitrov, David Jurgens, Andrew Piper, and Derek Ruths. 2016. Annotating characters in literary corpora: A scheme, the CHARLES tool, and an annotated novel. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 184–189, Portorož, Slovenia. European Language Resources Association (ELRA).

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Ralph Weischedel et al. 2013. Ontonotes release 5.0.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2008. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 176–183, Sapporo, Japan. Association for Computational Linguistics.

Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland. Association for Computational Linguistics.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2022. *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics, Gyeongju, Republic of Korea.

Zdeněk Žabokrtský and Maciej Ogrodniczuk, editors. 2023. *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*. Association for Computational Linguistics, Singapore.

Amir Zeldes. 2017. The gum corpus: Creating multilayer resources in the classroom. *Lang. Resour. Eval.*, 51(3):581–612.

Amir Zeldes. 2022. Opinion piece: Can we fix the scope for coreference? *Dialogue & Discourse*, 13(1):41–62.

Hongming Zhang, Hantian Ding, and Yangqiu Song. 2019a. SP-10K: A large-scale evaluation set for selectional preference acquisition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 722–731, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019b. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 872–881, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongming Zhang, Yan Song, Yangqiu Song, and Dong Yu. 2019c. Knowledge-aware pronoun coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 867–876, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2021. A brief survey and comparative study of recent development of pronoun coreference resolution in English. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–11, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. 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 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Xinran Zhao, Hongming Zhang, and Yangqiu Song. 2022. PCR4ALL: A comprehensive evaluation benchmark for pronoun coreference resolution in English. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5963–5973, Marseille, France. European Language Resources Association.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online. Association for Computational Linguistics.

A Limitations

We focus on a limited formulation of PCR. One could expand on the scope of these results by considering additional formulations of PCR as well as additional types of pronominal or other proform expressions (*e.g.,* a broader set of expressions considered as pronouns such as first and second person or reflexive pronouns). Additionally, the scope of coreference could be more explicitly specified by distinguishing identity coreference versus other related phenomenon such as binding.

We also did not consider differences between dataset annotation in detail. For example, datasets differ in the annotation of mention spans. Our experiments using gold mention annotations provide some insight into the impact of these differences, but this impact could be studied more thoroughly by explicitly considering how mention spans are annotated within each dataset.

Furthermore, we did not consider model failure cases in detail beyond our ablation experiments on the ARRAU dataset. For example, how performance might vary based on genre and how this differs between systems. For instance, LinkAppend was trained with genre and speaker metadata.

Finally, expanding our evaluations to multilingual pronominal anaphora and subsets of coreference datasets other than the English language would allow for new results regarding phenomenon that are more prominent outside of English (*e.g.,* zero-anaphors) or do not exist in English (*e.g.,* switch reference and obviation).

B Ethics Statement

PCR systems are known to perform disparately on subgroups which has ethical implications particularly for potential real-world use cases (Zhao et al., 2018; Rudinger et al., 2018; Webster et al., 2018; Hossain et al., 2023). We therefore do not recommend or endorse the use of these systems

for downstream purposes such as real-world, commercial applications; rather, our experiments are focused solely on the validity of certain assumptions of existing challenge sets.

C Differences in formulations of PCR

Our formulation of PCR follows the precise setup proposed by Zhang et al. (2021) which was in turn based on earlier formulations which also considered fixed subsets of English pronouns in restricted contexts. These restrictions were viewed as reasonable because most antecedents occur within the local context of a pronoun; e.g., Yang et al. (2003) observed that the antecedent is within the local context 95% of the time in the MUC corpus.

We similarly formatted WSC-like challenge sets in this way to allow for a fair comparison. For instance, WSC-like datasets may initially contain pronominal expressions outside our considered set such as *one* and *y'all*.

For additional details, we release our preprocessing code at `https://github.com/ianporada/challenge-set-assumption`.

C.1 Additional Considerations

In this section, we outline differences in how existing work has approached PCR and which choices we make in setting the scope of our analysis.

There is a tradeoff between evaluating all forms of pronominal coreference that might occur in natural language and evaluating those forms that have been identified and defined in such a way that they can be reliably annotated in existing corpora. With this perspective, our goal is more oriented towards the latter. That is, we do not intend to analyze all conceivable coreferences of all possible pronominal expressions. Rather, we take the intersection of existing work to better understand how well models generalize across datasets.

End-to-end v.s. mention-linking: As an end-to-end task, the goal of PCR is to determine with which linguistic expressions a pronoun corefers given only the raw context and identification of the pronoun. In contrast, it could be the case that candidate antecedent mentions are already identified, in which case the task of PCR consists of resolving the correct antecedent. Common approaches are to score each candidate independently or pairwise (Yang et al., 2008). We compare existing systems within the category with which they can perform the task.

One v.s. many mentions: A discourse entity can be realized as multiple coreferring linguistic expressions in a discourse. These realizations form a coreference cluster. In the case that multiple realizations appear in the context of a pronominal expression, there are multiple possible interpretations for what is considered the correct antecedent to be resolved to the pronoun. Popular approaches are to consider the most recent mention (Liang and Wu, 2003) or any one of the coreferring mentions as a valid antecedent. We consider the most recent mention to be the valid antecedent which is the approach most commonly taken in existing work (Zhang et al., 2021). Nonetheless, we do not consider instances where multiple coreferring expressions appear within the immediate context w to allow for our binary classification evaluation.

Mention boundaries: Finally, datasets differ in the annotation of mention boundaries (Moosavi et al., 2019). For example, the antecedent noun phrase "a young man" in Figure 1 is annotated in the dataset as "man" whereas in OntoNotes would be annotated as the maximal dominating span "a young man of large fortune from the north of England" according to the annotation guidelines. In the case where a PCR model functions end-to-end a reasonable assumption might be that the pronoun should corefer with at least a mention containing the head word of the correct antecedent and no mention containing the head word of the incorrect antecedent (Žabokrtský and Ogrodniczuk, 2022); however, optimizing for head words in this way has been shown to lead to strange modeling design decisions that do not align with human intuition (Žabokrtský and Ogrodniczuk, 2023). Moosavi et al. (2019) presents a method for normalizing mention boundaries so some minimal span which is a reasonable choice for end-to-end systems and may be useful for future work. For mention-linking, we simply consider the dataset's annotated mention. We do not consider more complex phenomenon such as split-antecedents and discontinuous mentions in our analysis, but these would also be interesting to investigate in future work.

D Input Format

In this section we provide more details regarding how the input was formatted. We also discuss additional results given other formats. We find that our conclusions are consistent across these decisions.

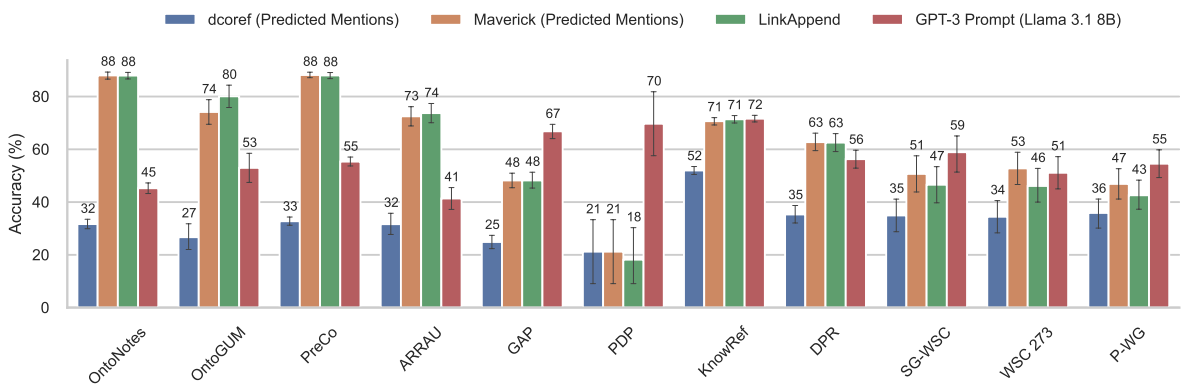

Figure 7: A comparison of the LinkAppend coreference resolution system and other systems that also do not rely on gold mention annotations. The GPT-3 Prompt is most accurate on certain WSC datasets, but performance is near random chance on OntoNotes, OntoGUM, PreCo, and ARRAU. This may be due to the difficulty in predicting the appropriate spans of linguistic expressions.

Speaker Information. Many datasets of coreference annotations consist of spoken language and include corresponding speaker metadata. We tested models both with and without this metadata and report models in their best configuration. We found that Maverick and dcoref perform best with speaker information, which including speaker data in LLM input in the form of "SPEAKER_NAME: ..." had marginal negative effect on LMs' performance.

Input Length. The datasets of curated natural corpora that we consider typically consist of relatively long document contexts. Therefore, we experimented with including in the input only the local context w or the entire document. We find that including the full context length has a marginal effect on performance. In the unsupervised case, we use only the local context w whereas for supervised models we include the full document in the input for all models (both finetuned systems and few-shot LMs).

dcoref For the dcoref baseline, we do not use gold parses (since not all datasets include parse information) and rather use parses predicted by the Stanford CoreNLP pipeline.

LinkAppend As an additional representative example of supervised systems, we consider the state-of-the-art LinkAppend coreference resolution system (Bohnet et al., 2023). We use the publicly released weights of the best performing system which consists of the multilingual mT5-XXL language model (13B params) finetuned on OntoNotes. Results for the LinkAppend system without gold mention spans are presented in Figure 7.

E Full Ensemble

Results of the ensembling method across all datasets are presented in Table 3. (We do not consider PDP due to its small test set size.) The ensembled predictions outperform any single system on all datasets except OntoNotes and KnowRef. The OntoNotes test set is known to have a high lexical overlap with the training set which could possibly explain the exceptional superior performance of the supervised Maverick model on this dataset. To test if this is the case, we can also consider the public Maverick weights trained on PreCo in the same setup in which case the Maverick model accuracy is 0.71 and is significantly outperformed by the ensemble approach (0.89 in this case).

In the case of KnowRef, it is not clear why the ensemble approach is less accurate than the supervised system in contrast to all other datasets. This may be related to the relatively poor accuracy of Llama 3.1 on KnowRef and would be interesting to investigate in future work.

F Dataset Details

F.1 Examples

Here we present example instances from the validation sets of those datasets that include a validation split. For readability, we show only the local context w.

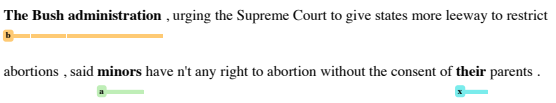

Figure 8: An instance from the OntoNotes dataset.

System	ON	OG	PreCo	ARRAU	GAP	KnowRef	DPR	SG-WSC	WSC 273	P-WG
Maverick	**0.96***	**0.90**	0.94	0.90	0.88	**0.87***	0.78	0.73	0.71	0.60
Llama 3.1 70B	0.87	0.84	0.88	0.87	0.77	0.73	0.82	0.77	0.82	0.80
Ensemble	0.93	**0.90**	**0.95**	**0.91**	**0.91***	0.77	**0.85**	**0.79**	**0.84**	**0.81**

Table 3: Accuracy of the ensemble method compared against Maverick (supervised coreference resolution) and prompted Llama 3.1 70B instruct. ON denotes OntoNotes and OG denotes OntoGUM. The most accurate model in each column is marked in bold. * indicates results that are statistically significant as compared to the next best model based on a t-test with p-value < 0.1.

In a representative democracy , however , the citizens do not govern directly . Instead , they elect representatives to make decisions and pass laws on behalf of all the people . Thus , **U.S. citizens** vote for members of **Congress** , the president and vice president , members of state legislatures , governors , mayors , and members of town councils and school boards to act on **their** behalf .

Figure 9: An instance from the OntoGUM dataset.

Balzac tells us of a man who suspected **his wife** of having a lover . The husband comes home by surprise . But she hears him and quickly hides **her lover** in the closet of **her** bedroom .

Figure 10: An instance from the PreCo dataset.

it is tempting for girls to try to hide their acne with make-up . This rarely hides the spots , and it blocks the skin pores - a situation almost guaranteed to make the acne worse . If **you** want to wear **make-up** , use **it** sparingly and choose a light non-greasy lotion , not cold creams .

Figure 11: An instance from the ARRAU dataset.

Coudert traveled to Europe in the late 1890s under the patronage of socialite Minnie Paget . She was welcomed by high society there and soon became known for her portraits of royalty , including **King Edward VII** , **Czar Nicholas II of Russia** and **his** wife Alexandra .

Figure 12: An instance from the GAP dataset.

William is often a condescending prick to **Frank** , but I do n't think **he** was being arrogant here .

Figure 13: An instance from the KnowRef-60K dataset.

Mike helped **Jack** with his assignment because **he** politely asked him to .

Figure 14: An instance from the DPR dataset.

Brian got a job working with dogs , while **Derrick** worked in sales because **he** was a people person .

Figure 15: An instance from the Pronominal Winogrande (P-WG) dataset.

F.2 Summary Statistics

OntoNotes

- License: LDC User Agreement for Non-Members
- Final validation instances: 1,536
- Final test instances: 1,642

OntoGUM

- License: Varies by subcorpus. All annotations are `cc-by-4.0`
- Final validation instances: 272
- Final test instances: 236

PreCo

- License: None specified
- Final validation instances: 2,167
- Final test instances: 2,248

ARRAU

- License: LDC User Agreement for Non-Members
- Final validation instances: 179
- Final test instances: 411

GAP

- License: `apache-2.0`
- Final validation instances: 203
- Final test instances: 832

PDP

- License: `cc-by-4.0`
- Final test instances: 33

KnowRef-60K

- License: `cc-by-4.0`
- Final validation instances: 21,240
- Final test instances: 3,061

DPR

- License: None specified
- Final test instances: 558

SuperGLUE WSC

- License: Custom (research usages)
- Final test instances: 146

WSC 273

- License: `cc-by-4.0`
- Final test instances: 180

Pronominal Winogrande

- License: `cc-by-4.0`
- Final test instances: 209