

Detecting Referring Expressions in Visually Grounded Dialogue with Autoregressive Language Models

Bram Willemsen and Gabriel Skantze

Division of Speech, Music and Hearing

KTH Royal Institute of Technology

Stockholm, Sweden

{bramw, skantze}@kth.se

Abstract

In this paper, we explore the use of a text-only, autoregressive language modeling approach for the extraction of referring expressions from visually grounded dialogue. More specifically, the aim is to investigate the extent to which the linguistic context alone can inform the detection of mentions that have a (visually perceivable) referent in the visual context of the conversation. To this end, we adapt a pretrained large language model (LLM) to perform a relatively course-grained annotation of mention spans in unfolding conversations by demarcating mention span boundaries in text via next-token prediction. Our findings indicate that even when using a moderately sized LLM, relatively small datasets, and parameter-efficient fine-tuning, a text-only approach can be effective, highlighting the relative importance of the linguistic context for this task. Nevertheless, we argue that the task represents an inherently multimodal problem and discuss limitations fundamental to unimodal approaches.

1 Introduction

In conversation, speakers often make reference to objects, events, or concepts. Words and phrases that are used with referential intent are known as *referring expressions* (REs). Effective communication hinges on the ability of the participants in the conversation to recognize these expressions and to determine what it is that they refer to, i.e., their *referents*. Within the context of a discourse, identification of an intended referent for a given RE may necessitate coreference resolution, i.e., the process of linking expressions that have the same referent. To illustrate this need, consider the following hypothetical exchange, with coreferring expressions underlined:

- (1) **A:** Have you seen Schrödinger’s cat?
- (2) **B:** Yeah, here it is.
- (3) **A:** It is looking a bit worse for wear.

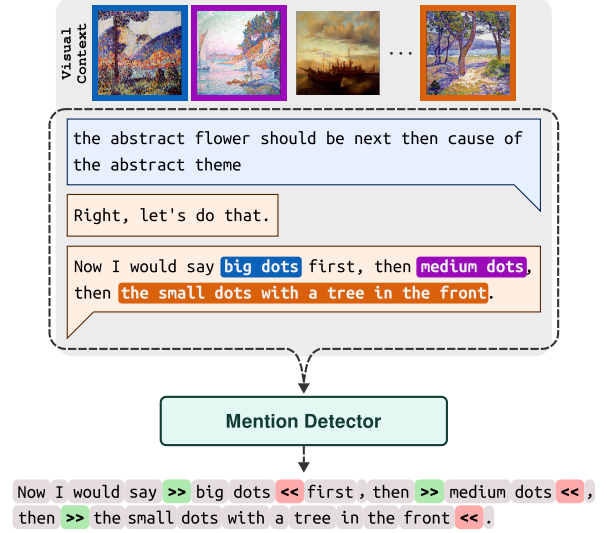


Figure 1: Visualization of the proposed mention detection method. The **Mention Detector** takes as input the most recent dialogue message, preceded by the available dialogue history, and autoregressively outputs an annotated reproduction of the last message while inserting mention span boundary tokens (the start and end of mention spans are represented by **>>** and **<<**, respectively) where appropriate. Excerpt shown is from a dialogue collected by Willemsen et al. (2022). Highlighted mentions in original dialogue and visual context with highlighted referent images are shown solely for illustrative purposes: the former is not available to the model at inference time, the latter neither at inference nor at training time.

Without access to the discourse context, “*it*” and “*It*” have indeterminate referents. By having knowledge of the prior contributions to the conversation, it is clear that both pronouns are anaphors with “*Schrödinger’s cat*” as their antecedent.

The identification of REs, or *mentions*¹, in various types of discourse is a long-standing natural language processing (NLP) task commonly referred to as *mention detection* (MD). Simply put,

¹We use *referring expression* and *mention* interchangeably throughout this paper.

when a given discourse is represented as a text document, the goal of MD is to identify any and all spans of text that refer to some predetermined type of referent, such as named entities or events.

In this paper, we explore the problem of MD for conversation, specifically with a focus on the downstream purpose of reference resolution in visually grounded dialogue. That is, our aim is to identify the REs that have a (visually perceivable) referent in the visual context of the conversation. Of particular interest is the extent to which the linguistic context alone is able to inform predictions for what is arguably, inherently, a multimodal problem. In addition, we experiment with different context windows to investigate how dialogue history affects MD performance. The expectation is that providing access to additional linguistic context in the form of preceding messages will generally lead to increased performance. To illustrate by example, whether the use of “*that*” in the exclaimed utterance “*Take that!*” is referential or instead part of a non-referential interjection cannot be known without additional context.

In line with recent work on generative information extraction (see e.g., Zhang et al., 2025), we frame MD in visually grounded dialogue as an autoregressive language modeling problem. More specifically, we propose to train a model to generate annotated reproductions of utterances: for a given utterance, in the process of generating a copy of the original message content, the model is expected to insert span boundary tokens indicating the start and end of mention spans, when and where appropriate. An illustration of the proposed approach is shown in Figure 1. Our experiments involve the parameter-efficient fine-tuning (Dettmers et al., 2023) of a large language model (LLM) on annotated conversations from two different visually grounded dialogue datasets, namely A GAME OF SORTS (AGOS, Willemsen et al., 2022) and PHOTOBOK (PB, Haber et al., 2019). For AGOS, we make use of the mention annotations from Willemsen et al. (2023). For PB, we adopt a similar annotation protocol to manually create the required mention annotations for a subset of the dataset.²

Results of our experiments with the 8B-parameter variant of LLAMA 3.1 (Grattafori et al., 2024) are promising, suggesting that the linguistic context can be relatively revealing for our purpose.

Note that our findings are in spite of the fact that our datasets are relatively small, our LLM is relatively moderately sized, and our fine-tuning is parameter-efficient. Nevertheless, we must contend with some limitations that are fundamental to unimodal approaches to multimodal problems, as well as the nature of the referential language in task-oriented dialogues. We provide additional discussion on these matters.

2 Background

MD has long been an essential component, or the central focus, of systems addressing various NLP tasks, such as named entity recognition (e.g., Lample et al., 2016; Devlin et al., 2019; Straková et al., 2019), event detection (e.g., Lai et al., 2020), and coreference resolution (e.g., Lee et al., 2013; Poesio et al., 2018). Earlier, rule-based approaches to MD were frequently built atop a dependency parse of a text, and would, over time, incorporate increasingly more powerful statistical models into the pipeline (e.g., Florian et al., 2010; Lee et al., 2013). The required sophistication of the approach generally depended on the downstream task. For coreference resolution, for example, simple heuristics leading to high recall would suffice if other parts of the system could compensate with higher precision (e.g., Lee et al., 2013). Interestingly, comparisons between different coreference resolution systems have often been conducted on the basis of gold, instead of predicted, mentions. This effectively side-steps MD in an effort to focus on isolating the system’s downstream performance. However, there tend to be notable performance gaps between these idealized and the realistic scenarios. As Poesio et al. (2023) note, generally, the overall performance of a coreference resolution system has been contingent on the accuracy of the output from its MD component.

Following advances in neural language modeling, approaches to MD based on neural models (e.g., Lample et al., 2016; Poesio et al., 2018; Devlin et al., 2019; Straková et al., 2019; Lai et al., 2020; Yu et al., 2020) have gradually superseded the earlier methods. These increasingly more data-driven methods promised to do away with the need for extensive feature engineering. Particularly consequential has been the adoption of general purpose, pretrained language models based on the Transformer architecture (Vaswani et al., 2017), examples of which include the encoder-only BERT

²<https://github.com/willemsenbram/mention-detection-vgd>, doi:10.5281/zenodo.15500581

(Devlin et al., 2019) and the decoder-only GPT (Radford et al., 2018). BERT-based representations have been the backbone of numerous NLP systems, including those that deal with MD (e.g., Devlin et al., 2019; Straková et al., 2019; Yu et al., 2020).

Of particular interest here are the autoregressive LLMs at the heart of most work on generative information extraction (see e.g., Zhang et al., 2025). Various studies have shown that framing tasks involving structured predictions as autoregressive language modeling problems can be effective (e.g., Cao et al., 2021; Liu et al., 2022; Deußner et al., 2024). Given an unstructured text, the model is trained to return, via next-token prediction, a structured representation of the input. Although the feasibility of this approach has been shown for commonly used benchmarks that involve some form of MD (e.g., Kim et al., 2003; Tjong Kim Sang and De Meulder, 2003), to the best of our knowledge, it has yet to be applied to visually grounded dialogue. In this paper, we explore to what extent we can adapt a pretrained LLM via parameter-efficient fine-tuning (Hu et al., 2022; Dettmers et al., 2023) to the task of MD in visually grounded dialogue using this approach.

3 Method

3.1 Problem description

In general, the goal of MD is to identify all expressions in a document D that satisfy some prescribed definition of a *mention*. When D is a *visually grounded* dialogue, we define it as $D = (V, L)$, where V is the visual context and L the linguistic context of the conversation. A dialogue is considered visually grounded when L contains one or more references to V . That is, within the linguistic context, there exists one or more expressions that have a (visually perceivable) referent that is present in the visual context of the conversation.

3.2 Task definition

In this work, we consider MD in visually grounded dialogue to be the task of identifying all expressions in L for which there exists a referent in V . Here, we focus on visually grounded dialogues of which V is composed of a set of v independent images, $V = \{I_1, I_2, \dots, I_v\}$. The linguistic context L can be represented as a sequence of n utterances³, $L = (u_1, u_2, \dots, u_n)$. In turn, each utterance u_i

³We use *utterance* and *message* interchangeably throughout this paper.

can be represented as a sequence of m_i tokens, $u_i = (t_{i1}, t_{i2}, \dots, t_{im_i})$. We think of mentions in terms of spans. We can define a mention span as a contiguous subsequence of tokens from an utterance u_i , denoted as $(t_{ij}, \dots, t_{ik}) \subseteq u_i$, where $1 \leq j \leq k \leq m_i$. Together, these tokens constitute an expression that (indirectly) refers to one or more of the images. Note that in contrast with other types of documents, dialogue is interactive and contributions to L are cumulative, happening over time. It is important to account for the incremental nature of conversation when addressing this task.

3.3 Proposed approach

Core to our approach is the framing of MD in visually grounded dialogue as a next-token prediction task. Given the incremental nature of conversation, we process each dialogue at the utterance level, prepending to each utterance a token indicating the speaker. For a given utterance u_i , we train an autoregressive language model f to reproduce exactly the original content of u_i , but with span boundary tokens inserted if and where appropriate to indicate the start and end of mention spans.

Crucially, however, we propose to condition the generation of the target sequence u_i' not only on the current utterance u_i , but also on additional preceding linguistic context, i.e., the available dialogue history, as prior messages may inform predictions. When considering prior messages in the modeling process, we can define the generation of u_i' as $u_i' = f(u_i, H)$, where H is the dialogue history available to the model.⁴ The available dialogue history H is defined as a contiguous subsequence of utterances from L , denoted as $H = (u_{i-h}, u_{i-h+1}, \dots, u_{i-1})$, where $0 \leq h \leq w$, where h is the number of prior messages available to the model and w is an optionally predefined maximum number of preceding messages to be considered. For a visualization of the proposed approach, see Figure 1.

4 Experiments

The language modeling experiments presented in this paper involve the fine-tuning of pretrained models on dialogues from two different, though

⁴We must note that for the experiments reported in this paper, we found that repeating utterance u_i in the input to the model had a positive impact on downstream performance; a slight deviation from the more general definition provided here. For an example of the formatting of training samples, see Appendix B.

closely related, visually grounded dialogue tasks, namely A GAME OF SORTS (AGOS, [Willemssen et al., 2022](#)) and PHOTOBOOK (PB, [Haber et al., 2019](#)). We first perform cross-validation to score MD performance on each dataset separately. We then assess cross-dataset transfer by training on one dataset and testing on the other. In addition, we investigate the effects of dialogue history on MD performance, i.e., whether the model benefits from having access to preceding messages when making its predictions, by experimenting with different context window sizes, i.e., providing access to different numbers of preceding messages. Finally, as points of comparison, we assess the MD performance of a baseline based on noun phrase (NP) extraction using constituency parsing, as well as that of an encoder-only LLM fine-tuned for sequence labeling.

4.1 Data

Both AGOS and PB are tasks designed around eliciting repeated references to various sets of real-world images—such as those found in the MS COCO ([Lin et al., 2014](#)) and Open Images ([Kuznetsova et al., 2020](#)) datasets—in conversational settings. Moreover, both tasks have a deliberate asymmetry in their visual contexts that participants have to overcome to successfully complete the task. This ensured that speakers would produce non-trivial REs that made reference to the images’ visual content.

4.1.1 A Game Of Sorts (AGOS)

AGOS is a collaborative image ranking task. Two participants are shown a set of nine images which they are asked to rank, in descending order and one at a time, based on a given sorting criterion. The goal of the task is for the participants to, through conversation, arrive at a ranking which both deem satisfactory. Although both participants see the same set of images, they cannot see each other’s perspective. The position of the images on their respective screens has been randomized, forcing the participants to refer to the images by referencing their visual content. To ensure repeated mentions of the same referents, the task is performed over multiple (four) rounds, and the same set of images is used each round.

The AGOS dataset consists of 15 dialogues. Each AGOS image set consists of nine images from the same of one of five image categories, namely cars, dogs, paintings, pastries, or phones.

	AGOS	PB-GOLD
# Dialogues	15	50
# Messages (\circ)	1,800	3,335
# Mentions (✎)	1,486	2,111
# Characters (A)	86,516	96,774
# Words (AB)	19,843	22,889
% \circ with ✎	60.33%	61.02%
% \circ with > 1 ✎	17.94%	1.95%
# A in ✎	27,574	61,771
% A in ✎ : A in \circ	31.87%	63.83%
# AB in ✎	5,708	12,880
% AB in ✎ : AB in \circ	28.77%	56.27%
\bar{X} A in \circ	48.06 (43.57)	29.02 (24.83)
\bar{X} A in ✎	18.56 (15.76)	29.26 (23.35)
\bar{X} AB in \circ	11.02 (9.52)	6.86 (5.40)
\bar{X} AB in ✎	3.84 (3.20)	6.10 (4.86)

Table 1: Descriptive statistics for the AGOS and PB-GOLD datasets. *Note.* Explanation of symbols and abbreviations: \circ = Messages; ✎ = Mentions; **A** = Characters; **AB** = Words; \bar{X} = average (mean). Standard deviation between brackets. Scores and standard deviations are rounded to the nearest hundredth.

Three dialogues were collected per image category.

4.1.2 PhotoBook (PB)

PB is a collaborative image identification task. Two participants are shown partially dissimilar sets of six visually similar images; some of the images will be shown to both participants, while others are shown to only one of the participants. Each participant has three of their six images highlighted. The goal of the task is for the participants to, through conversation and without seeing each other’s perspective, identify for these highlighted images whether or not they have them in common. To ensure repeated mentions of the same referents, the task is performed over multiple (five) rounds, and while the set of images shown to participants changes from round to round, the image sets are constructed in such a way that each image is shown multiple times to at least one of the participants.

The PB dataset consists of 2.5K dialogues. Each PB image set, as shown to each participant, consists of six images that prominently feature two objects, each object belonging to a different image category. These two image categories form the “image domain” of the conversation; each image shown throughout the interaction will feature at least one object from each category. For our experiments, we make use of the so-called PB-GOLD subset, as referenced in [Takmaz et al. \(2022\)](#), which

		AGOS				PB-GOLD			
		0	3	7	19	0	3	7	19
LLAMA	P	.896 (.03)	.922 (.02)	.919 (.02)	<u>.923</u> (.03)	.933 (.02)	.936 (.03)	.940 (.02)	<u>.943</u> (.02)
	R	.835 (.04)	.865 (.03)	.883 (.03)	<u>.884</u> (.03)	.927 (.01)	.925 (.01)	.934 (.01)	<u>.937</u> (.02)
	F_1	.863 (.02)	.892 (.01)	.900 (.01)	<u>.902</u> (.01)	.930 (.01)	.930 (.02)	.937 (.02)	<u>.940</u> (.02)
	J	.811 (.03)	.849 (.03)	.856 (.02)	<u>.858</u> (.02)	.921 (.01)	.922 (.01)	<u>.933</u> (.01)	<u>.933</u> (.01)
M-BERT	P	.827 (.04)	.842 (.03)	.843 (.03)	<u>.863</u> (.04)	.916 (.02)	.918 (.02)	.924 (.01)	<u>.930</u> (.02)
	R	.812 (.05)	.835 (.03)	.837 (.04)	<u>.853</u> (.01)	.909 (.01)	.912 (.01)	.908 (.01)	<u>.917</u> (.02)
	F_1	.819 (.04)	.838 (.02)	.839 (.02)	<u>.857</u> (.02)	.912 (.02)	.915 (.01)	.916 (.01)	<u>.924</u> (.02)
	J	.786 (.04)	.815 (.02)	.814 (.02)	<u>.825</u> (.01)	.909 (.01)	.914 (.01)	.913 (.01)	<u>.920</u> (.01)

Table 2: Cross-validated mention detection performance of fine-tuned LLAMA 3.1 8B (LLAMA, top) and MODERNBERT-large (M-BERT, bottom) on AGOS and PB-GOLD for four different context windows, i.e., 0, 3, 7, and 19 preceding messages. *Note.* P = Precision; R = Recall; F_1 = F_1 score; J = Jaccard index. Scores are rounded to the nearest thousand, standard deviations to the nearest hundredth.

consists of 50 dialogues for which the authors have provided some annotations at the utterance level.

4.1.3 Mention annotations

In this work, we make use of the manually annotated mention spans from Willemsen et al. (2023). These spans indicate the linguistic expressions that have a (visually perceivable) referent in the visual context of the conversation. More specifically, these are either singletons or REs that are part of an identity relation with other mentions in the linguistic context that have one or more of the images as their referents. For the annotation of the mention spans, Willemsen et al. (2023) were aided by speakers’ self-annotations, as participants were required to indicate whether or not a message was meant to include one or more references to one or more of the images. In the messages which contained such references, the longest, most specific spans with images as their referents were marked. The resulting annotations are relatively course-grained. We adopt this protocol for our annotation of the PB-GOLD dialogues. Although PB has no self-annotations, referential ambiguities can be resolved by scrutiny of the full dialogue context. We report descriptive statistics of both datasets in Table 1.

4.2 Model specifications

For each experiment involving the proposed autoregressive language modeling approach, we fine-tune LLAMA 3.1 8B (Grattafiori et al., 2024) using QLoRA (Detmers et al., 2023) on a single 24GB NVIDIA GeForce RTX 3090. We calculate the loss only over tokens of the target message, masking the loss over tokens that are part of the preceding dialogue context. We make use of the model’s existing vocabulary for any special tokens, such as those

indicating span boundaries. Fine-tuned model output is generated using constrained decoding. That is, at every time step we dynamically restrict the vocabulary, where the allowed tokens include the next token from the input utterance and any valid special tokens. Hyperparameters are listed in Table 6 in Appendix A. For an example of the formatting of training samples for fine-tuning, see Appendix B. For additional implementation details, we refer the reader to our repository.²

4.2.1 Baselines

NP extraction using constituency parsing As mentions are predominantly NPs, we opt for a simple baseline model that automatically extracts NPs from the dialogues using the constituency parser from the Stanza toolkit (Qi et al., 2020). The backbone of this parser is ELECTRA-large (Clark et al., 2020) trained on a revised version of the third release of the Penn Treebank (Marcus et al., 1993). We extract the most expansive spans, but discard certain candidate phrases. For instance, as the dialogues involve text-based conversations in which the participants are not able to see each other, we can disregard various personal pronouns (e.g., “I”, “you”, “me”) as these were not considered to be mentions here.

Sequence labeling with MODERNBERT It has been common practice to treat problems that center on the detection of spans in text, such as MD, as sequence labeling tasks (e.g., Lample et al., 2016). When given a sequence (of tokens), the objective is to assign each element a label such that span boundaries can be inferred. Tag sets are frequently based on the IOB format (Ramshaw and Marcus, 1995): the B tag indicates that an element begins a span, the I tag indicates that an element is inside of a span,

and the O tag indicates that an element is outside of a span. For our purpose, we adopt the IOB tag set and fine-tune MODERNBERT-large (Warner et al., 2024) to predict for each token of a given utterance the correct label. As the name suggests, MODERNBERT is a more recent encoder-only LLM that improves upon the original BERT architecture. Similar to the LLAMA-based experiments, for the experiments that are meant to demonstrate the effects of dialogue history on downstream performance, we provide preceding messages as context, masking the loss over all labels except those of the target message. Each model is fine-tuned on a single 24GB NVIDIA GeForce RTX 3090. Hyperparameters are listed in Table 7 in Appendix A. For additional implementation details, we refer the reader to our repository.²

Note that in this formulation of the problem using the basic IOB format, it is not possible to accurately label nested mentions. However, there are very few cases of nesting in the datasets used for the experiments reported in this paper. Therefore, this shortcoming has negligible impact on the current evaluation of the approach.

4.3 Evaluation

Our first experiments involve cross-validation on both datasets. We evaluate using the same five-fold cross-validation protocol adopted by prior work on the AGOS dataset (Willemssen et al., 2023; Willemssen and Skantze, 2024), which partitions the dataset along its five image sets. We similarly perform five-fold cross-validation on the PB-GOLD dataset. However, as there is no predefined, deterministic split for PB-GOLD, we split the data randomly. Our second set of experiments concerns an investigation into cross-dataset transfer. This means that we fine-tune models on the entirety of AGOS and test on the entirety of PB-GOLD, and vice versa.

In addition, we test the effects of dialogue history on MD performance. For each of the aforementioned experiments, we fine-tune models for four different context windows, 0, 3, 7, and 19, meaning the models have access to no, three, seven, or 19 preceding messages, respectively.

4.3.1 Metrics

We measure mention detection performance in terms of precision, recall, F_1 score, and intersection over union of ground truth (gold spans) and predicted mention spans at the character level (i.e.,

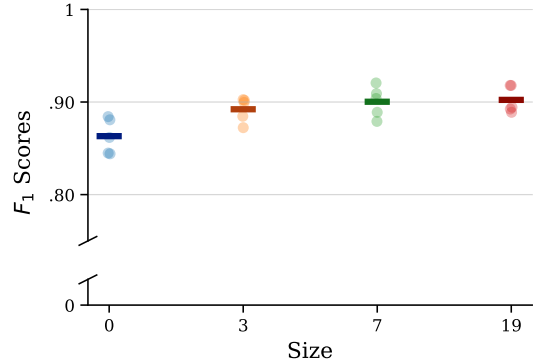


Figure 2: Mention detection performance of fine-tuned LLAMA 3.1 8B in terms of F_1 scores $[0, 1]$ as a function of the size of the context window, i.e., the maximum number of preceding messages considered from the available dialogue history. Shown are results of each fold (dots) and their average (bar) for four different context windows, i.e., 0, 3, 7, and 19.

Jaccard index).⁵

We calculate precision, recall, and F_1 scores based on exact mention span matches. This means that a predicted mention is considered a true positive only if it matches a gold span exactly and is treated as a false positive otherwise. Conversely, a ground truth mention for which there is no exact matching prediction is considered a false negative.

We use a measure based on the Jaccard index to score the extent to which ground truth and predicted mention spans overlap, which permits the scoring of partial matches. For each message, we find the optimal assignment of predicted and ground truth spans based on the number of corresponding character indices. We calculate the Jaccard index for each pair of matched spans. In the event that no match exists—that is, there is no overlap between a ground truth mention and any of the predicted spans (false negative), or there exists no ground truth mention for a predicted span (false positive)—, the score for this particular span is 0.

All the aforementioned mention detection metrics are bound $[0, 1]$, with higher scores indicating better performance.

5 Results

Before reporting the results of our fine-tuning experiments, we first highlight some of the descriptive statistics reported in Table 1 to aid in understanding the composition of the data. As shown in Table 1, PB-GOLD contains over three times

⁵Character-level evaluation avoids tokenization issues when span boundary tokens are placed within words.











						\bar{X}
P	.881	.954	.934	.912	.933	.923 (.03)
R	.897	.842	.902	.924	.855	.884 (.03)
F_1	.889	.894	.918	.918	.892	.902 (.01)
J	.853	.823	.881	.874	.857	.858 (.02)

Table 3: Cross-validated mention detection performance of fine-tuned LLAMA 3.1 8B on AGOS based on a context window of 19, i.e., a dialogue history consisting of 19 preceding messages. Results are shown for each fold as well as their average (\bar{X}). *Note.* P = Precision; R = Recall; $F_1 = F_1$ score; J = Jaccard index; Symbols represent folds:  = Cars;  = Dogs;  = Paintings;  = Pastries;  = Phones. Standard deviation between brackets. Scores are rounded to the nearest thousand, standard deviations to the nearest hundredth.

more dialogues than AGOS. However, on average, the AGOS dialogues are considerably longer and have almost twice as many messages per dialogue. While the percentage of messages with mentions is comparable, AGOS has a much higher rate of messages that contain more than one mention than PB-GOLD. Nevertheless, mentions make up notably less of the overall content of the AGOS dialogues than of the PB-GOLD dialogues; the number of characters and words dedicated to mentions relative to the total number of characters and words in the messages is substantially lower for AGOS than for PB-GOLD. Finally, the average AGOS mention is shorter than the average PB-GOLD mention.

5.1 Cross-validation

Shown in Table 2 are the cross-validated results from fine-tuning and evaluating models on the AGOS and PB-GOLD datasets. For each context window, scores are reported as averages over all folds for each MD performance metric. In addition, the results reported in Table 3 are from fine-tuning and evaluating LLAMA on the AGOS dataset using the maximum context window size we considered for this work, i.e., a context window of size 19. In Table 3, scores are shown per fold in addition to their averages over all folds, for each MD performance metric. We found that, despite some variance between folds, scores resulting from fine-tuning LLAMA were relatively high overall. In comparison, the performance of MODERNBERT is relatively competitive, but it does lag behind that of LLAMA. The observed results suggest that the models were, on average, somewhat more performant on the PB-GOLD than they were on the AGOS data. Moreover, we observed that the mod-

els generally benefited from an increase in context window size; on average, we found that providing the models with a greater number of preceding messages increased MD performance, but noted that there were diminishing returns. The observed trend was somewhat more apparent for AGOS than for PB-GOLD. Figure 2 provides a visualization of this trend based on the F_1 scores for AGOS.

5.2 Cross-dataset transfer

Table 4 shows results from fine-tuning models on AGOS and testing on PB-GOLD (AGOS \rightarrow PB-GOLD), and vice versa (PB-GOLD \rightarrow AGOS). Although scores were shown to trail those of the cross-validation experiments, the observed MD performance was still indicative of a relatively high degree of successful transfer overall. Again, LLAMA’s performance was shown to exceed that of MODERNBERT. A noteworthy observation was that AGOS \rightarrow PB-GOLD consistently resulted in higher scores than PB-GOLD \rightarrow AGOS on all MD performance metrics. Similarly to results from our cross-validation experiments, we observed that, on average, an increase in the size of the context window tended to result in improved performance. These findings suggest that providing the models with at least some preceding messages can already be beneficial.

5.3 Comparison with NP extraction

The results reported in Table 5 show the MD performance of a method based on constituency parsing for the automatic extraction of NPs. Although recall may seem relatively high considering that the focus of this baseline model was solely on NP extraction, it bears repeating that most mentions tend to be NPs, though they are not always presented in a straightforward, parsable manner or context. Perhaps unsurprisingly, especially when comparing against our proposed approach, this naive method for MD is relatively imprecise, as the false positive rate ends up being relatively high when predicting virtually all NPs to be referential in nature.

5.4 Error analysis

When examining the output generated by LLAMA, we found various errors to be consistent between the different context windows. Although the models appeared to be relatively robust against the noise in the input, certain mentions were partially, or entirely, missed, as a result of ungrammatical phrasing. For partial matches, we observed some

		AGOS → PB-GOLD				PB-GOLD → AGOS			
		0	3	7	19	0	3	7	19
LLAMA	P	.798	.859	.864	<u>.886</u>	.775	.803	.810	<u>.820</u>
	R	.806	.838	<u>.858</u>	.845	.687	.676	.713	<u>.744</u>
	F_1	.802	.848	.861	<u>.865</u>	.728	.734	.758	<u>.780</u>
	J	.777	.816	.834	<u>.839</u>	.668	.666	.694	<u>.722</u>
M-BERT	P	.725	.768	.778	<u>.795</u>	.707	.735	.774	<u>.777</u>
	R	.650	.694	.687	<u>.735</u>	.610	.641	.704	<u>.723</u>
	F_1	.685	.729	.730	<u>.764</u>	.655	.685	.737	<u>.749</u>
	J	.662	.704	.698	<u>.737</u>	.595	.616	.665	<u>.688</u>

Table 4: Mention detection performance of fine-tuned LLAMA 3.1 8B (LLAMA, top) and MODERNBERT-large (M-BERT, bottom) in cross-data transfer experiments for four different context windows, i.e., 0, 3, 7, and 19 preceding messages. **AGOS → PB-GOLD** indicates training on AGOS and testing on PB-GOLD; **PB-GOLD → AGOS** indicates training on PB-GOLD and testing on AGOS. *Note.* P = Precision; R = Recall; $F_1 = F_1$ score; J = Jaccard index. Scores are rounded to the nearest thousand, standard deviations to the nearest hundredth.

	AGOS	PB-GOLD
P	.411	.377
R	.764	.607
F_1	.535	.465
J	.453	.530

Table 5: Mention detection performance of the Stanza NP extraction baseline. *Note.* P = Precision; R = Recall; $F_1 = F_1$ score; J = Jaccard index. Standard deviation between brackets. Scores are rounded to the nearest thousand, standard deviations to the nearest hundredth.

recurring errors in relation to structural ambiguities, leading to the exclusion of relative clauses or prepositional phrases, and the splitting of single into multiple mentions or the merging of multiple mentions into a single span. Furthermore, we found instances of ambiguous pronoun usage to be relatively frequent among errors, such as in the phrases “*let’s go for it*” and “*let’s do it*”, in which the use of “*it*” is referential, but it is not recognized as such without additional context. Interestingly, providing access to preceding messages ends up resolving the inaccuracy for the former and not for the latter, even though these seem to be very similar cases on the surface. Conversely, we also observed cases where usage of (pro)nouns was incorrectly predicted to be referential. Again, some of these errors were resolved by providing the model access to the dialogue history.

6 Discussion

In this paper, we explored the potential of an approach to mention detection (MD) in visually grounded dialogue based on autoregressive language modeling. Results from our experiments

on conversations from the visually grounded dialogue tasks A GAME OF SORTS (AGOS, Willemssen et al., 2022) and PHOTOBOOK (PB, Haber et al., 2019) were promising, showing that a text-only approach that involves the parameter-efficient fine-tuning of LLMs to generate annotated reproductions of utterances can be effective. Moreover, we showed that providing the models with additional context from the dialogue history—that is, any messages that preceded the utterance under consideration—generally benefits performance. Although these findings were largely consistent between the competing methods presented in this work, within our experimental setup the generative approach to information extraction using the fine-tuned, decoder-only LLAMA model was shown to consistently outperform the sequence labeling approach based on the fine-tuned, encoder-only MODERNBERT.

Results from our cross-validation experiments showed that the models, on average, achieved better performance on the PB-GOLD than on the AGOS dataset. The cross-dataset transfer experiments revealed a notable performance gap between the two datasets; fine-tuning on the AGOS data seemed to result in the models being better able to generalize beyond their specific conversational domain than when fine-tuning on the PB-GOLD data. These findings suggest that AGOS offers a more challenging testbed when it comes to MD, as it was explored in this work, than PB-GOLD. Given that the primary focus of the PB task is the correct identification of images, participants’ language use is disproportionately reserved for referential purposes. This was made apparent through a quantified characterization of the PB-GOLD mentions, indicating

that mentions made up nearly two-thirds of the linguistic content of the dialogues. In contrast, with image identification being a secondary objective, mentions make up just shy of one-third of the linguistic content of the AGOS dialogues. In addition, mentions in the PB-GOLD dialogues are considerably longer, on average, than those in the AGOS dialogues. When qualitatively examining the mentions from both datasets, it becomes clear that the incidence rate of mentions that resemble image caption-like descriptions is notably higher for PB-GOLD than for AGOS. By and large, our findings suggest that AGOS offers its referring language use in a richer linguistic context than PB-GOLD, which aids the models’ ability to generalize.

That being said, it would be reasonable to assume that the incidence rate of mentions in these task-oriented dialogues from both datasets is high compared to that of organic, non-task-oriented conversations. Conversations can go long stretches of time without the mention of a visually perceivable referent. Our approach relies heavily on there being exploitable regularities in the linguistic context. The extent to which conversations with comparatively sparse mention occurrences, and that take place outside of task-oriented settings, still exhibit such actionable patterns is, as of yet, unclear. For both AGOS and PB-GOLD, the probability that a given linguistic expression (indirectly) points to a referent that is visually perceivable by at least one of the participants in the conversation is high, simply as a consequence of the situational context, as the images are the focal point of the conversations. Discerning, from the linguistic context alone, whether an RE has such a referent becomes far more challenging, if not impossible, when the configuration of the visual context of the conversation is less constrained, more dynamic, and cannot be anticipated ahead of time. In other words, we may still be able to extract mention candidates with a high degree of accuracy, but the number of false positives—by which we here mean any candidates that currently have no visually perceivable referents—is likely to be significantly higher; this outcome reminds of the high recall settings favored by aforementioned prior work on coreference resolution.

Inevitably, a general solution to the problem will require a cross-modal approach. Although we make no assumptions regarding the manner of encoding, the visual information must somehow be incorporated to validate whether candidate men-

tions indeed have a referent in the visual context; even when the linguistic context strongly implies the existence of such a referent, we simply cannot be certain without a review of the visual context. Moreover, we are likely to see that end-to-end approaches will increasingly be favored over modular systems when it comes to addressing downstream tasks that have historically relied on some form of MD, but for which MD is simply a means to an end. Nevertheless, we expect that MD as a task in and of itself will remain relevant for niche applications for the foreseeable future. For one, it may continue to serve as a benchmark for the information extraction capabilities of models under varying conditions. Perhaps more interestingly, however, are real-world applications, such as its use as an information extraction tool for corpus linguistics.

Limitations

In this work, the focus has been on detecting REs that have a (visually perceivable) referent in the visual context of a conversation. Only singletons and mentions in an identity relation were considered, contingent on their referent being one or more of the images in the visual context. It is worth noting that there are some consequential differences between the images used by AGOS and PB. Where the focus of each AGOS image was on (an iconic view of) one entity from some image category, PB images depicted more complex scenes, purposely featuring multiple entities from different image categories. Perhaps unsurprisingly, when the task involves identification within a visually grounded conversational context, we find that the more complex the scene, the more frequently we have to consider a bridging relationship between mentions as a surrogate for identity. This highlights a complication with respect to the annotation of this domain that becomes increasingly problematic: the noisier (or more complex) the language use, the more ambiguous the boundaries. We expect this to be even more evident in unrestricted, spoken dialogue.

Regarding our cross-validation experiments, results were based on a five-fold split of the datasets. The AGOS dataset has a preferred partitioning that ensures minimal data leakage between the training and test data. For PB-GOLD, however, we did not find a sensible, deterministic split, as even when image domains were seemingly mutually exclusive, in reality there were frequent intrusions from other image categories. For instance, *people*—

which happens to be one of the author-defined image categories—are present in the vast majority of the photographs, often as salient entities, and frequently referenced as a result. Although we do not believe this has affected our overall conclusions, the random splitting may have resulted in inflated scores in the PB-GOLD cross-validation experiments. In addition, the language used in the dialogues from both datasets is exclusively English, meaning the experiments reported in this paper do not provide explicit insight into the extent to which the approach generalizes to other languages.

Finally, we have evaluated the proposed approach with one LLM undergoing a parameter-efficient fine-tuning regimen. We have not investigated performance differences between full-parameter and parameter-efficient fine-tuning, nor have we tested the extent to which other generative LLMs are able to perform the task. In addition, more exhaustive hyperparameter tuning has the potential to improve results further. It is conceivable that more optimal hyperparameters exist that could narrow the observed performance gap between LLAMA and MODERNBERT on this task. However, it would mainly serve to underscore the general importance of the linguistic context and demonstrate the viability of either approach.

Acknowledgements

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The authors would like to thank Dmytro Kalpakchi, Jim O'Regan, Travis Wiltshire, Chris Emmery, Martina Rossi, and the anonymous reviewers for their helpful comments.

References

- Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. [Autoregressive Entity Retrieval](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient Fine-tuning of Quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.
- Tobias Deußner, Lars Hillebrand, Christian Bauckhage, and Rafet Sifa. 2024. [Informed Named Entity Recognition Decoding for Generative Language Models](#). In *2024 IEEE International Conference on Big Data (BigData)*, pages 6001–6010.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Radu Florian, John Pitrelli, Salim Roukos, and Imed Zitouni. 2010. [Improving Mention Detection Robustness to Noisy Input](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 335–345, Cambridge, MA. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The Llama 3 Herd of Models](#). *_eprint*: 2407.21783.
- Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. [The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-Rank Adaptation of Large Language Models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.
- Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. [GENIA corpus - a semantically annotated corpus for bio-textmining](#). In *Proceedings of the Eleventh International Conference on Intelligent Systems for Molecular Biology*, pages 180–182, Brisbane, Australia.
- Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. [The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale](#). *International Journal of Computer Vision*, 128(7):1956–1981.

- Viet Dac Lai, Tuan Ngo Nguyen, and Thien Huu Nguyen. 2020. [Event Detection: Gate Diversity and Syntactic Importance Scores for Graph Convolution Neural Networks](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5405–5411, Online. Association for Computational Linguistics.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural Architectures for Named Entity Recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- 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.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: Common Objects in Context](#). In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.
- Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. [Autoregressive Structured Prediction with Language Models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. [Building a Large Annotated Corpus of English: The Penn Treebank](#). *Computational Linguistics*, 19(2):313–330.
- Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjatz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. [Anaphora Resolution with the ARRAU Corpus](#). In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, Louisiana. 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.
- 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, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. [Improving language understanding by generative pre-training](#).
- Lance Ramshaw and Mitch Marcus. 1995. [Text Chunking using Transformation-Based Learning](#). In *Proceedings of the Third Workshop on Very Large Corpora*, pages 82–94.
- Jana Straková, Milan Straka, and Jan Hajic. 2019. [Neural Architectures for Nested NER through Linearization](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5326–5331, Florence, Italy. Association for Computational Linguistics.
- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. [Less Descriptive yet Discriminative: Quantifying the Properties of Multimodal Referring Utterances via CLIP](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.
- Erik F. Tjong Kim Sang and Fien De Meulder. 2003. [Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition](#). In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is All you Need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. [Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference](#). *_eprint*: 2412.13663.
- Bram Willemsen, Dmytro Kalpakchi, and Gabriel Skantze. 2022. [Collecting Visually-Grounded Dialogue with A Game Of Sorts](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2257–2268, Marseille, France. European Language Resources Association.
- Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. [Resolving References in Visually-Grounded Dialogue via Text Generation](#). In *Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–469, Prague, Czechia. Association for Computational Linguistics.
- Bram Willemsen and Gabriel Skantze. 2024. [Referring Expression Generation in Visually Grounded Dialogue with Discourse-aware Comprehension Guiding](#). In *Proceedings of the 17th International Natural Language Generation Conference*, pages 453–469, Tokyo, Japan. Association for Computational Linguistics.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. [Neural Mention Detection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1–10, Marseille, France. European Language Resources Association.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025. [A Survey of Generative Information Extraction](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

A Hyperparameters

As a starting point for hyperparameter optimization, we took note of hyperparameters reported in prior work (e.g., [Hu et al., 2022](#); [Dettmers et al., 2023](#); [Warner et al., 2024](#)), performing minimal tuning mostly within suggested ranges.

Epochs	2
Batch size	8
Learning rate (LR)	1e-4
LR scheduler type	cosine
Warmup ratio	0.1
LoRA r	16
LoRA α	16
LoRA dropout	0
LoRA target modules	*_proj, lm_head

Table 6: Hyperparameters for QLoRA fine-tuning of LLAMA 3.1 8B. We use default values if not otherwise specified.

Epochs	4
Batch size	8
Learning rate	8e-5
Gradient accumulation steps	8
Warmup ratio	0.1
Weight decay	8e-6

Table 7: Hyperparameters for fine-tuning of MODERN-BERT-large. We use default values if not otherwise specified.

B Training example

The following is an example of a training sample from the AGOS dataset—for a context window of size 3—that was used to fine-tune LLAMA 3.1:

B: Clear, I think my second choice would be the light grey one, which looks like in old style.\nA: I agree, its bottom seems to be pretty high as well.\nB: yeap!\nB: then, for the third one, is the dark grey one okay?\n\nB: then, for the third one,

is the dark grey one okay? -> B: then, for the third one, is >> the dark grey << one okay?

Messages in the linguistic context are separated by single newline characters (`\n`). Each message is prepended with a token indicating the speaker (either **A** or **B**). The message we want annotated is separated from the linguistic context by two newline characters (`\n\n`). This message is followed by an inference token (`->`). The inference token is then followed by the annotated message, with span boundary tokens indicating the start (`>>`) and end (`<<`) of the mention span.