Do *they* mean 'us'? Interpreting Referring Expressions in Intergroup Bias

Anonymous ACL submission

Abstract

The variations between in-group and out-group speech (intergroup bias) are subtle and could underlie many social phenomena like stereotype perpetuation and implicit bias. In this paper, we model the intergroup bias as a *tagging task* on English sports comments from forums dedicated to fandom for NFL teams. We curate a unique dataset of over 6 million game-time comments from opposing perspectives (the teams in the game), each comment grounded in a non-linguistic description of the events that precipitated these comments (live win probabilities for each team). Expert and crowd annotations justify modeling the bias through tagging of implicit and explicit referring expressions and reveal the rich, contextual understanding of language and the world required for this task. For large-scale analysis of intergroup variation, we use LLMs for automated tagging, and discover that some LLMs perform best when prompted with *linguistic descriptions* of the win probability at the time of the comment, rather than numerical probability. Further, large-scale tagging of comments using LLMs uncovers **linear variations in the form of referent** across win probabilities that distinguish in-group and out-group utterances.

1 Introduction

Social bias in language is generally studied by identifying undesirable language use towards a specific demographic group (Kaneko and Bollegala, 2019; Sheng et al., 2019; Sap et al., 2020; Webson et al., 2020; Pryzant et al., 2020; Sheng et al., 2020); However, we can enrich our understanding of bias in communication by understanding it as differences in behavior situated in social relationships. Intergroup bias is the social bias stemming from the intergroup relationship between the speaker and target reference of an utterance. (Maass et al., 1989; Maass, 1999). Govindarajan et al. (2023a) modeled intergroup relationships (in-group and out-group) and interpersonal emotions in interpersonal

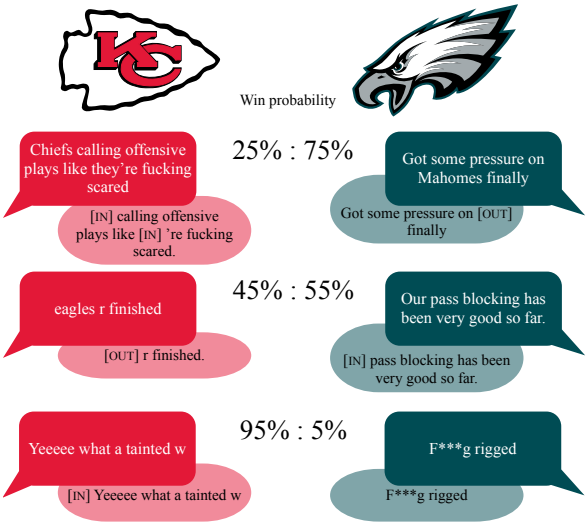

Figure 1: We construct a parallel language corpus of comments from NFL team subreddits, grounding each comment in the live win probabilities. We then tag relevant entities in each comment with intergroup tags using annotators and LLMs.

English language tweets at the utterance level, finding systematic interactions between these two parameters. While neural models based on LLMs can be trained to discriminate in-group and out-group utterances, causal probing of these models was inconclusive (Govindarajan et al., 2023b) and still left major questions unanswered:

- How does language systematically change when referring to an individual in one's in-group versus their out-group?
- How does the state-of-the-world influence this systematic variation?

In this work, we take major steps towards answering these questions, using a task architecture in the classical NLP pipeline (Manning et al., 2014). Earlier work in the Linguistic Intergroup Bias (LIB) hypothesis (Maass, 1999) focused exclusively on the predicate, and the bias was described using an ad-hoc lexical categorization system (Semin and Fiedler, 1988). However, the **form of referencing** the in-group or out-group can reveal subtle

biases as well. Consider the comments in Figure 1, taken from the dataset we describe in this paper. Commenters refer to the in-group and out-group by name, sub-groups, pronouns as well as implicitly — sometimes they choose not to refer to either group at all. How does the intergroup bias manifest in referent forms?

To answer this question, we introduce a new dataset of interpersonal language — comments from game threads on online forums dedicated to fandoms for teams in the National Football League (NFL). Through careful data curation, we construct a parallel corpus of sports comments, with comments from fans of both teams in a game, *aligned* in time and *grounded* in win probabilities (WP). By focusing on referring expressions, we can formulate investigating the intergroup bias as a **tagging** task: given a comment, the group affiliation of the writer, and the state-of-the-world, return a *tagged comment* with appropriate referring expressions tagged as [IN], [OUT] or [OTHER] (see Figure 1). Annotation and preliminary analysis reveal that the form of the referent that speakers use when referring may have systematic intergroup variations.

We train Large Language Models (LLMs) to automate large-scale tagging of our dataset, and examine their performance on our task. We find that few-shot performance on GPT-4o is boosted using linguistic descriptions of win probabilities; fine-tuned Llama-3 models performed better, although incorporating WP had little effect. Using our best performing model to tag 100,000 comments from our raw dataset, we discover two striking linguistic behaviors:

1. Higher the win probability for the *in-group*, the more likely commenters are to **abstract away** from referring to the in-group. This trend is remarkably linear across win probabilities for all types of in-group references.
2. References to out-groups by commenters are rarer than in-group references, and remain stable over all win probabilities for the in-group.

These findings add much needed color to the LIB hypothesis — natural language is productive, and commenters can express their (implicit) intergroup bias in different ways. This work also lays the foundation for future explorations of other intergroup variations (in event descriptions, for example) in sports-talk and other domains. We share all our code, models, and data online.

2 Background and Related Work

Intergroup bias Linguistic Intergroup Bias (LIB) theory (Maass et al., 1989; Maass, 1999) hypothesizes that stereotypes are transmitted and persist in communication through systematic linguistic asymmetry — socially desirable in-group behaviors and socially undesirable out-group behaviors are encoded at a higher level of abstraction. The LIB has been reproduced in psychological experiments and analyses (Anolli et al., 2006; Gorham, 2006); it has also been used as an indicator for a speaker's prejudicial attitudes (Hippel et al., 1997), and racism (Schnake and Ruscher, 1998).

Govindarajan et al. (2023a,b) take inspiration from the LIB at large to study intergroup bias as a general phenomenon in online language use. While they find regularities in its variation with emotion that neural models can 'learn' to identify in-group and out-group utterances more accurately than humans, probing experiments fail to describe human-observable intergroup variations in language. This work studies a much larger dataset than in their work, and by modeling the bias as a tagging task to referents, we discover characteristic lexical variations at scale that complement LIB findings.

Sports language Language use in the domain of sports has been a rich source of analyses and studies within computational linguistics, including from the perspective of quantifying *social biases*. Merullo et al. (2019) studied commentator racial biases in descriptions of football players, reaffirming previous findings illustrating clear differences in terms of sentiment descriptions (white players were more likely to be described as intelligent), and name itself (white players were more likely to be referred to by their first name). Zhang et al. (2019) focused on one aspect of language usage among (and between) fans of NBA teams: intra-group behavior with and without social contact with the out-group. They find that fans with intergroup contact are more likely to use negative language — they were more polarized than before.

Our work differs from previous work in two major ways. Firstly, we focus on the intergroup bias — how do fans talk about their team (in-group), versus the opponent (out-group)? Secondly, this paper **grounds** the analysis of intergroup bias in numerical descriptions of the state-of-the-world. The state-of-the-world in a sports game at any moment can be described using the scoreboard, thus

providing grounding for utterances follow. Non-linguistic, numerical descriptions of the events that precipitate an utterance overcome the drawbacks of using ad-hoc, proximally derived metrics like *social desirability* (in LIB) or *affect* (in Govindarajan et al. (2023b)) as an axis to study linguistic variation. As we shall describe in §3, sports games, and in particular NFL games, are rich with statistical information amenable to describe the state-of-the-world on a well-calibrated numerical scale.

3 Data & Annotation

3.1 Dataset

Data & Preprocessing Our new dataset of intergroup language comes from Reddit — specifically subreddits dedicated to fandoms for each of the 32 teams in the NFL. During the NFL season, each subreddit has *game threads* — posts created by moderators on which fans can comment in tandem with the live game involving their team. Crucially, since every subreddit has their own thread, we effectively have a **parallel intergroup language dataset**; two teams and their fans commenting on *the same game events*. Further, these subreddits are dedicated to individual team fandoms, so we can fairly assume that the team the subreddit represents is the in-group for all commenters.[1]

We focus on all completed games from the 2021–22 and 2022–23 NFL seasons, and attempted to scrape all comments from the game threads for both teams involved in every game. Within comments from game threads, we filtered it down to comments that happened during active game-time, and removed comments that were only URL links. Overall, our raw data has over 6 million comments from 768 game threads on 32 subreddits, grounded in 491 NFL games.

Grounding football comments American Football has some attractive features as a sport considering that our interest is in the *language surrounding the events* in a game — it is highly strategic, and outcomes are heavily dependent on a coach's strategies and plays in a (relatively) small number of discrete events (called *plays,* Pelechrinis and Papalexakis, 2016). The state-of-the-world at any moment in a football game is determined by a variety of factors — seconds remaining in half (and game), yard line, score differential, down, yards to go, home advantage, timeouts remaining, betting odds lines from Vegas, and so many more (Horowitz et al., 2017; Yurko et al., 2018). Baldwin (2021) modeled the **Win Probability** (henceforth **WP**) of a team at any point during the game using a decision trees over the aforementioned features, building a well-calibrated model with low error. We chose WP as a succinct, non-linguistic description of the events preceding an utterance.

Using the `nflFastR` (Baldwin, 2021) package, we can obtain WPs for individual plays in each game, as well as the time of completion of a play. Combined with the timestamps at which the comments were submitted (obtained from the Reddit API), we build our parallel corpus of intergroup language grounded in win probabilities. The WP cleverly models the complexities of a real-world sporting event into one number that accurately models how **desirable** the state-of-the-world is to the in-group (see Figure 1).

3.2 Tagging

As we motivated in §1 and Figure 1, tagging references to entities enables us to perform analyses at scale *and* discover individual lexical variations. Consider the following examples:

(1) a. Rams are gifting us a chance to win and we can't take advantage. The f***!!!!
b. if the ravens and chiefs beat these dudes by double digits then damn it so should we!

Even without contextual information about the game for the above comments, we see *multiple* readily identifiable references to the in-group and out-group, within the same utterance. The words or phrases that refer to relevant individuals can now be tagged with in-group ([IN]) or out-group ([OUT]) For instance, (1) would be tagged thus:

(2) a. [OUT] are gifting [IN] a chance to win and [IN] can't take advantage. The f***!!!!
b. if [OTHER] and [OTHER] beat [OUT] by double digits then damn it so should [IN]!

We define the in-group ([IN]) as the team the commenter supports (and its fans), and the out-group ([OUT]) as the opponent in that particular game (and its fans). The spans 'the ravens' and 'chiefs' in (1-b) are clearly not a reference to the in-group nor the opponent of the game. However, they are a reference to *a group of interest in this domain* — another NFL team and/or its fans. We consider these references to be [OTHER], and a special case of out-group references.

[1] Note that we focus on language of online commenters (fans) on Reddit, not *commentators* for the game.

Sometimes, the references to the in-group, out-group or other are not explicit. However, we can infer based on common-sense reasoning that the comment as whole, or a sentence in the comment, is **implicitly referring** to a relevant group:

(3) What a conservative play call

There is no explicit word/phrasal reference to any team in the above comment. However, it is clear in context (the fan's team is losing, with WP of 9%) that the commenter is referring to the in-group. To facilitate these implicit annotations, we sentence tokenize the comments in our dataset using Stanza (Qi et al., 2020), append a sentence-level token [SENT] before each sentence in every comment in our dataset. If the sentence as a whole is judged to implicitly refer to a relevant group, the [SENT] token is replaced with the relevant tag.

3.3 Annotation

Annotators are presented with a comment from our dataset, the source subreddit (team) for the comment, the parent comment (if the comment is a reply in a thread), and the live score at the time of the comment. The task of tagging words and phrases from comments in our dataset with intergroup tags can be highly involved — in addition to knowledge of American Football, commonsense reasoning over the meaning of an utterance in context of the live game, one needs knowledge of the teams and its players. For instance, in (4), one needs to know that the commenter supports the Seahawks, and that there is a prominent player named Wilson, to accurately tag in context that Wilson indeed is an in-group reference.

(4) Our oline should start holding since apparently it 's okay now . Maybe Wilson can actually get some time to throw .

Implicit annotations on the [SENT] token require a higher bar of reference, since all comments are about the game at hand and will involve both teams to some extent. For example, we judge the following comments to not have explicit or implicit references to any relevant groups of interest even though they are about the game:

(5) a. Fair enough !
b. winning cures all lmao
c. turning the game off , have a good day yall

In case it is impossible to verify an explicit or implicit reference, annotators are instructed to not highlight any parts of the comment. All annotators were free to search the web for names or expressions they were unfamiliar with, as well as refer to reports of the game to understand the utterance completely, and accurately tag all references. All annotation experiments were carried out using the `thresh.tools` annotation interface (Heineman et al., 2023). Annotators highlight spans within a comment and select from one of 3 tags, and select a confidence level from a five-point scale.

Expert annotated dataset We gather expert annotations for constructing a 'gold' annotated dataset to evaluate crowd annotations and modeling moving forward. The first author of the paper annotated 1499 comments (randomly sampled from all game-time comments) for intergroup references based on a pre-defined, written protocol (described in detail in Appendix B). 26.7% of comments were judged to have no relevant intergroup reference, and in the remaining comments, references to the in-group (76.3%) vastly out-number references to the out-group (14.6%) or other. This is not surprising, since these are comments from forums dedicated to fandom of teams — people are much more likely to talk about their team over the opponent. We partition our gold dataset into a test set of 318 datapoints, and a training set of 1181 datapoints.

Crowd annotation To understand our dataset further, we recruited three undergraduates to annotate the test split of our expert dataset. Our goals were to understand where disagreements arose, as well as how and when knowledge of the events in the game helped in disambiguating references. Annotators were given similar instructions, and were free to search the web and lookup statistics and reports on the game in question. We found in pilot experiments that the live-score was more interpretable to humans than WP, influencing our choice to provide that as context to annotators.

4 Preliminary Analysis

4.1 Annotation Analysis

Inter-annotator agreement Average Fleiss κ (Fleiss, 1971) among crowd annotators is 0.69, indicating moderate agreement. In addition to the inter-annotator score, by counting exact-matches and weighting partial matches between individual crowd annotators and gold annotations, we calculate an 'accuracy' score of 0.65 ± 0.005. This can be interpreted as a human ceiling for performance on this task, and characterizes its inherent subjectivity and difficulty.

Disagreement can be a signal Looking at the source of disagreements among annotators (and between crowd annotators and experts) can give us insights into the nature of the task itself (Atwell et al., 2022), as well as why differences in judgements of intergroup affiliation can come down to annotator biases or judgement given context. For example, annotators disagree sometimes on what counts as a 'reference':

(6) a. . . . **Lambeau** has the second worst bathrooms .

b. Can't do that against **an offense** this good.

Lambeau in (6-a) was judged by the expert annotator to be a reference to the out-group in this context — the opponent/out-group is the Green Bay Packers. However, to tag this referent [OUT], annotators would need to know or deduce that Lambeau Field is the Green Bay Packers stadium, and judge that this constitutes a relevant intergroup reference. Thus, disambiguating some references can be time-consuming and hard. *an offense* in (6-b) was judged by some annotators to refer to the out-group in context. However the generic nature of the referent lead other annotators to judge that this was an overall statement about the game, rather than an explicit reference.

Whether or not examples in (6) contain references to the in/out-group is not simply a consequence of the difficulty of our task, or the inability for annotators to transparently describe the mental state of commenters. Rather, we need to analyze them as possibly another subtle influence of the intergroup bias itself — demonstrated by questioning why commenters chose the forms in (6) rather than in (7), which convey the same meaning, and would be uncontroversial in annotation:

(7) a. . . . **the Packer's** stadium has the second worst bathrooms .

b. Can't do that against **a Packers** offense. . .

4.2 Qualitative Analysis & Trends

Mereology of referring expressions Expert annotation revealed that commenters refer to groups of interest in a myriad of different ways. In the previous section, we liberally defined the annotation protocol for highlighting references to *individuals* in the in-group, out-group and other. Using insights from mereology (Varzi, 2019), we derive a taxonomy of parthood in intergroup relations, that defines what it means for a reference to constitute a reference towards the in-group/out-group/other:

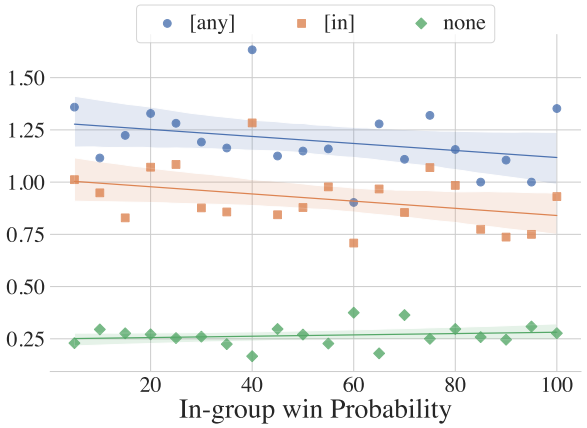

Figure 2: Per-comment frequency of in-group, any and 'none' references in gold dataset over WP.

1. **People**: Names, nicknames, shirt numbers, initials, pronouns, etc. : *Tua, TK87, he/him. . .*
2. **Subset of the team**: This refers to groups of players, or coachers, rather than just one player: *the offense*, *our defense*, *o-line*, . . .
3. **Team**: Name of the team (*rams, bills, cowboys*), nicknames (*lambs, cowgirls*), city names(*LA, Buffalo, Dallas*), pronominal expressions like *our boys* for the in-group, pronouns like *they/them* for the in-group and out-group, and many more.
4. **Team plus supporters**: The first person pronouns *we* and *us*, but can also be done with the third person pronouns *they* and *them*. The latter of course, could also refer to out-group or other, and require context to disambiguate.

The taxonomy above is ordered in order of increasing coverage of the whole group, by the referring part — the size of the reference gets larger from people to the entire group. Thus, players are the smallest unit of reference within a group, and the team/organization plus its supporters constitute the largest possible reference to the group itself.

Trends The annotated dataset enables us to study qualitative trends, that will guide quantitative modeling analyses presented in §6. We specifically focus on two phenomenon that are directly observable in the data and illustrated with examples — diversity in form of referring expression, and trends over WP. Within the gold dataset, we can observe two clear trends by plotting the frequency of a feature of interest over comments that fall within a win probability (WP) window. Figure 2 plots the frequency of any reference, in-group references, and 'None' references over all 5% WP windows:

1. References to the in-group, and references to

any group overall, go down with WP.

2. 'None' references increase steadily with WP.

The steady increase in number of 'None' references in higher WP windows is interesting, but requires robust analysis. While the trends observed in this section are not statistically significant, this can be attributed to the small sample size of only 1499 comments. The intergroup bias is a social phenomenon, and like many social phenomenon, we can make clear inferences at scale. Obtaining human annotated data at scale would be prohibitively hard and expensive in this setting — we use LLMs, to automate this task, thus allowing us make inferences about trends in the intergroup bias as a function of WP.

5 Modeling intergroup bias with LLMs

Large Language Models (LLMs) have shown remarkable abilities in various domains over the last few years (Srivastava et al., 2023; Brown et al., 2020). Our novel tagging framework to model intergroup bias requires linguistic understanding, knowledge of the NFL and its teams, as well as complex reasoning over why a commenter might choose certain word forms compatible with the state-of-the-world — making LLMs well suited to this task. In this section, we design modeling experiments to tag comments from our dataset with intergroup labels towards two objectives:

- Understand how LLMs statistically 'reason' over meaning in context of an utterance and game state (WP) to tag comments.
- Discover hidden intergroup variations in referring expressions by tagging a large sample of comments from our raw, scraped data.

5.1 Modeling conditions

We focus on two specific models — Llama-3-8B (AI@Meta, 2024) and GPT-4o (OpenAI, 2024). Both are decoder based models that perform best at a wide variety of benchmarks, and allow us to compare and contrast the performance of an open-weights model with finetuning, versus a larger closed model with few-shot prompting. Building upon previous work, we prompt both models with a combination of instructions, chain-of-thought explanations, and few-shot examples (Wei et al., 2022). Llama-3 is prompted with the same input format, but we also finetune the model on the train split of our gold dataset. See Appendix D for further details on training and inference.

CoT Explanations We finetune Llama-3 with GPT-4o generated CoT explanations (Wadhwa et al., 2023). We first generate a explanation from GPT-4o for each comment in our gold dataset using instructions, few-shot examples, the target tagged comment and list of referring expressions provided as input to GPT-4o. All few-shot explanations were written by the first author, and examples were drawn from outside the gold dataset.

Our task is framed end-to-end as the model receiving the untagged comment as input with some contextual information (in-group, out-group, WP), and being asked to generate the comment with relevant words/phrases replaced with the appropriate tags. To understand the impact of WP on model performance we design 3 conditions

Numeric WP The model receives WP as a numeric input — a percentage between 0 and 100 that is WP for the in-group.

No WP WP is not provided as input to the model, and the instructions nor few-shot explanations neither use nor mention it.

Linguistic WP We experiment with providing WP as a scalar description of game state, from 'Team A is very likely to win' to 'Team B is very likely to win' based on the numeric WP corresponding to the comment.

We also experimented with utilizing the WP to modify the temperature when decoding (Atwell et al., 2022). When temperature scaling (TS) is used, we set the temperature to $sin(\pi.WP)$ — this pushes the LM to choose less likely words when the game's outcome is more uncertain.

Evaluation To evaluate the performance of a model on the test dataset, we report **micro-F1** scores for each of the three tags, and a weighted macro-F1 score overall. To give partial credit for the model's tagged output slightly overlapping with the gold tagged spans, we assign partial scores (0.5 and 0.25) for being within 3 and 5 characters of the correct tagged spans respectively.

5.2 Results

Table 1 shows the results for both of our models on all conditions. While both models exceed the human baseline performance that we calculated in §4, Llama-3 nudges GPT-4o overall. GPT-4o performs better at identifying out-group and other references by names or nicknames due to its much larger size and more parametric knowledge.

	Model	Random Baseline	Numeric	Numeric WP+TS	No WP	No WP +TS	Ling. WP +TS	Ling WP
GPT-4o	[IN]	35.6(3.2)	66.6(1.4)	67.4(2.0)	67.7(1.9)	69.9(1.8)	71.2(0.7)	**71.7(0.8)**
	[OUT]	20.1(1.1)	64.6(3.1)	**67.1(2.0)**	63.6(2.7)	66.6(1.0)	63.4(1.9)	63.7(2.5)
	[OTHER]	14.0(5.9)	**54.1(1.1)**	53.9(2.1)	49.0(1.6)	47.5(2.1)	48.4(4.7)	51.6(8.8)
	Overall	30.8(2.7)	64.9(1.3)	65.9(1.5)	65.0(1.4)	66.9(1.3)	67.4(1.3)	**68.2(0.3)**
Llama-3-8b	[IN]	35.6(3.2)	72.0(1.5)	72.0(1.5)	72.5(0.6)	72.1(0.4)	72.6(1.9)	**72.6(1.9)**
	[OUT]	20.1(1.1)	**60.2(2.3)**	57.9(4.3)	59.9(0.7)	58.3(4.3)	58.7(3.5)	57.7(4.7)
	[OTHER]	14.0(5.9)	59.2(8.8)	58.8(8.2)	**64.0(4.2)**	57.6(6.8)	59.8(4.5)	59.1(5.4)
	Overall	30.8(2.7)	68.8(2.4)	68.4(2.0)	**69.6(0.9)**	68.4(1.0)	68.9(2.5)	68.8(2.6)

Table 1: Results from few-shot experiments on GPT-4o (top), and finetuning Llama-3-8b (bottom).

WP helps... sometimes? Including WP did not change the performance of Llama-3 noticeably. As we observed in annotation, there are few examples of comments being ambiguous enough that the state-of-the-world is enough to disambiguate what a reference could be. Entire classes of references (from our taxonomy in §4.2) are quite unambiguous even without whole-sentence context.

We do observe however that providing WP in language form boosts the in-group tagging performance of GPT-4o in few-shot settings, besting numeric WPs ($p < 0.005$ with a bootstrap test). Analysis of model's outputs reveal GPT-4o's fickleness and inability to reason over numerical scales — for instance it reasons that WPs ranging from 1% to as high as 41% are 'low' in its explanations. Further, it rarely uses the numbers to infer the WP for the out-group in explanations. Since we re-write low WPs with the name of the out-group (as winning) in the linguistic WP condition, this might explain the model's slight boost in performance.

While Llama-3's performance is better through finetuning, it does not benefit from incorporating WP in training or inference. We attempted scaling the loss during training with WP and expert annotator confidence ratings where available, but these didn't boost performance. Whether larger LLMs exhibit similar behaviors to GPT-4o when finetuned, we leave to future work.

Error analysis While phrasing the WP with words improves tagging performance on in-group referents (over numerical WP as Table 1 shows), especially with GPT-4o, performance on out-group references remains stable. However, we do observe the model making similar 'errors' like annotators that we described in §4.1 — for instance, GPT-4o occasionally tagged *a single WR* (an indefinite/generic entity) as out-group in (8-a), which was not judged to be a relevant referent in expert annotation. However, GPT-4o also makes some basic errors, such as tagging *a catch* as out-group in (8-b) as well.

(8) a. Wait wait wait. Did I hear that right? **They** don't have **a single WR** with a catch today??
b. Wait wait wait .Did I hear that right ? [OUT] don't have [OUT] with [OUT]??

6 Analysis of model-tagged comments

Our novel tagging framework is amenable to application on a large sample of our raw data, facilitating us to observe and analyze variations in how members refer to different groups as a function of WP. We sample 100,000 comments from a larger raw dataset, and apply our best performing finetuned Llama-3 model towards this task. Since WP could not be effectively incorporated to improve performance, we used the model finetuned with no WP information. Further, we verified there was no correlation between the model's accuracy and WP — accuracy mostly followed comment density across WP (see Appendix A). After inference from our finetuned LLM, we use regular expressions to ensure that any obvious words (names and nicknames of teams, *we/us*) were also tagged appropriately, and to count different referring expressions accounting for inflections.

Figure 3 plots the frequency of different references over WP. We divide WP into 5% windows, and count the number of comments that contain a specific tag (or tag-lexical item pair), and divide it by the number of comments within that window in the entire sample. Figure 4 plots a few more reference variables of interest and is similar except the variables are normalized by number of comments that contain *any* reference. There are two findings we wish to highlight.

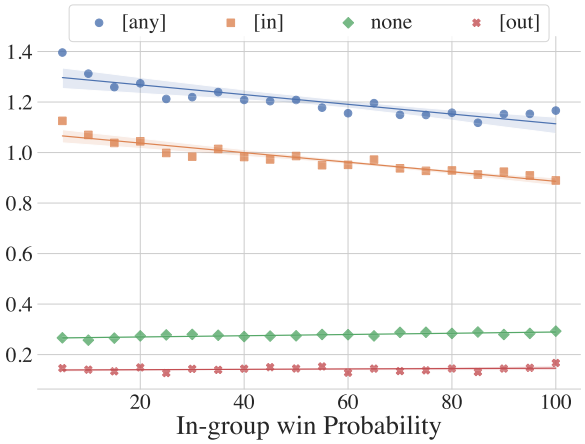

Figure 3: Per-comment frequency of various reference variables over all 5% WP windows. A 95% CI regression line is fit separately for each variable.

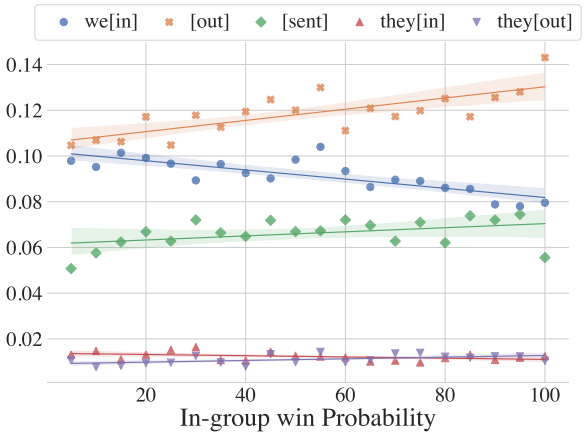

Figure 4: Normalized per-comment frequency of various reference variables over all 5% WP windows. A 95% CI regression line is fit separately for each variable.

Winning trumps all Figure 3 clearly shows a linear, decreasing tendency for commenters to refer to any entity or the in-group the more likely the team is to win. Observing a sample of high WP (9)) comments reveals an increase in positive excitement, but also increased terseness closer to victory:

(9) a. HOLY S**T
b. WHAT A THROW

Figure 4 illustrates the tendency to refer less to the in-group, is compounded with an increased tendency to refer to the out-group, or to refer implicitly over a sentence (the [SENT] lemma) when referring at all over WP. Overall, this paints a clear picture — the more likely the in-group is to win, commenters prefer to refer to the out-group or to refer to entities implicitly if they refer at all; They prefer to abstract away from specific events and express excitement the closer the in-group is to winning.

WP as a well calibrated predictor A striking feature of Figures 3 and 4 is the linear relationships between reference variables and WP. Table 2 in Appendix E estimates coefficients and R-squared for linear fits, but we can observe visually that with increasing WP, commenters are more likely to refer to the out-group with 'they' than the in-group. From Figure 4, the slope for in-group references ($-2.8e^{-}4$) is larger than the slope for references to the in-group using first person singular pronouns ($-2e-4$); Commenters are more likely to refer to in-group using the most inclusive term at higher WPs, when referring to the in-group at all.

These findings add to the subtle ways we perpetuate bias in our linguistic behavior, especially towards **in-group protection** (Maass, 1999). While commenters are more than willing to criticize the in-group across WP, the self-protective instinct is evident in the way they choose to refer to the in-group using *we/us* more often when losing, the reduced tendency to refer to the in-group using *they/them*, or to not refer to the in-group at all. Thus, how commenters choose the form of reference to an in-group constitutes just as subtle a bias as their choice of predicate.

7 Conclusion

We expand the study of intergroup bias by building a parallel corpus of sports comments grounded in win probabilities from live games. Annotation experiments reveal that modeling the bias as a tagging problem over words can reveal unobserved variations, as well as make it amenable to large-scale modeling. Through few-shot and finetuning experiments, we find that LLMs can out-match human baseline performance at this task, but struggle to reason over win probabilities, or use it meaningfully towards tagging. Tagging a large sample of our dataset reveals linear trends between various referring expressions and WP, showing that intergroup bias can manifest in commenter's choice of who to refer to when commenting on a game and how. Careful data curation and understanding, combined with focused usage of LLMs as statistical information processing tools can thus reveal linguistic variations in social language use online at scale. In future work we plan to exploit the parallel nature of our corpus further to understand team differences in language variation, as well as how WP can be effectively incorporated into a model of social meaning.

Limitations

Our work expands the study of intergroup bias in language by focusing on natural language use in online conversations on the Reddit platform. Further, our focus on grounding the utterances lead us to focus on sports talk, specifically conversations around NFL games. Biases in demographics of users on Reddit, or demographics of NFL fans are thus inherent in our data and analysis. Future work needs to study the prevalence of our findings in other sports with similar statistics that enables efficient grounding of utterances, as well as in more general speech.

We identify that both few-shot performance by GPT-4o and finetuned performance by Llama-3 are close to, or out-perform the human ceiling performance. Human ceiling performance is simply the average accuracy of crowd annotators against expert annotators. As we note in the paper, this is a difficult and inherently subjective task. Our results do not mean that models (finetuned or not) have a better understanding of what constitutes intergroup references, nor that they are more aligned with the task. Llama-3 was trained on the training split of the expert annotated gold data-set. While GPT-4o was exposed to the same set of examples as human annotators, it is a very large (possibly a mixture of trillions of parameters) model that contains a multitude of statistical associations that aids in instruction following.

Ethics

We downloaded comments from Reddit threads using the official Reddit API, and will disseminate our data in accordance with the Reddit terms of service. We will only release the comment and post ids for the raw data, and usernames will be anonymized. We will release the annotated data in full with the same precautions. We have censored some of the profanity in the comments when used as examples in this paper, since our focus isn't on abusive/negative language exclusively.

All created artifacts from this work (code, annotated data) will be released under the MIT License.

Crowd-sourced annotations were collected from three undergraduates employed by one of the authors for 15$ an hour.

References

AI@Meta. 2024. Llama 3.

Luigi Anolli, Valentino Zurloni, and Giuseppe Riva. 2006. Linguistic Intergroup Bias in Political Communication. *The Journal of General Psychology*, 133:237 – 255.

Katherine Atwell, Remi Choi, Junyi Jessy Li, and Malihe Alikhani. 2022. The role of context and uncertainty in shallow discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 797–811, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ben Baldwin. 2021. Open source football: nflfastr ep, wp, cp xyac, and xpass models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Bradley W. Gorham. 2006. News Media's Relationship With Stereotyping: The Linguistic Intergroup Bias in Response to Crime News. *Journal of Communication*, 56(2):289–308. Place: United Kingdom Publisher: Blackwell Publishing.

Venkata Subrahmanyan Govindarajan, Katherine Atwell, Barea Sinno, Malihe Alikhani, David I. Beaver, and Junyi Jessy Li. 2023a. How people talk about each other: Modeling generalized intergroup bias and emotion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2496–2506, Dubrovnik, Croatia. Association for Computational Linguistics.

Venkata Subrahmanyan Govindarajan, David Beaver, Kyle Mahowald, and Junyi Jessy Li. 2023b. Counterfactual probing for the influence of affect and specificity on intergroup bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12853–12862, Toronto, Canada. Association for Computational Linguistics.

David Heineman, Yao Dou, and Wei Xu. 2023. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 336–345, Singapore. Association for Computational Linguistics.

W. Hippel, Denise Sekaquaptewa, and P. Vargas. 1997. The Linguistic Intergroup Bias As an Implicit Indicator of Prejudice. *Journal of Experimental Social Psychology*, 33:490–509.

Maksim Horowitz, Ron Yurko, and S Ventura. 2017. nflscrapr: Compiling the nfl play-by-play api for easy use in r.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving Debiasing for Pre-trained Word Embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Anne Maass. 1999. Linguistic Intergroup Bias: Stereotype Perpetuation Through Language. In Mark P. Zanna, editor, *Advances in Experimental Social Psychology*, volume 31, pages 79–121. Academic Press.

Anne Maass, Daniel Anthony Salvi, Luciano Arcuri, and Gün R. Semin. 1989. Language use in intergroup contexts: the linguistic intergroup bias. *Journal of personality and social psychology*, 57 6:981–93.

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.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. 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 6355–6361, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2024. Hello gpt-4o.

Konstantinos Pelechrinis and Evangelos Papalexakis. 2016. The anatomy of american football: evidence from 7 years of nfl game data. *PLoS one*, 11(12):e0168716.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically Neutralizing Subjective Bias in Text. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):480–489.

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*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Sherry B Schnake and Janet B Ruscher. 1998. Modern racism as a predictor of the linguistic intergroup bias. *Journal of Language and Social Psychology*, 17(4):484–491.

G. R. Semin and K. Fiedler. 1988. The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, 54:558–568. Publisher: American Psychological Association.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The Woman Worked as a Babysitter: On Biases in Language Generation. 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 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Achille Varzi. 2019. Mereology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Spring 2019 edition. Metaphysics Research Lab, Stanford University.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2023. Revisiting relation extraction in the era of large language models. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2023:15566–15589.

Albert Webson, Zhizhong Chen, Carsten Eickhoff, and Ellie Pavlick. 2020. Are "Undocumented Workers" the Same as "Illegal Aliens"? Disentangling Denotation and Connotation in Vector Spaces. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4090–4105, Online. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In

Advances in Neural Information Processing Systems, volume 35, pages 24824–24837. Curran Associates, Inc.

Ronald Yurko, Samuel L. Ventura, and Maksim Horowitz. 2018. nflwar: a reproducible method for offensive player evaluation in football. *Journal of Quantitative Analysis in Sports*, 15:163 – 183.

Jason Shuo Zhang, Chenhao Tan, and Qin Lv. 2019. Intergroup contact in the wild: Characterizing language differences between intergroup and single-group members in nba-related discussion forums. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

A Data

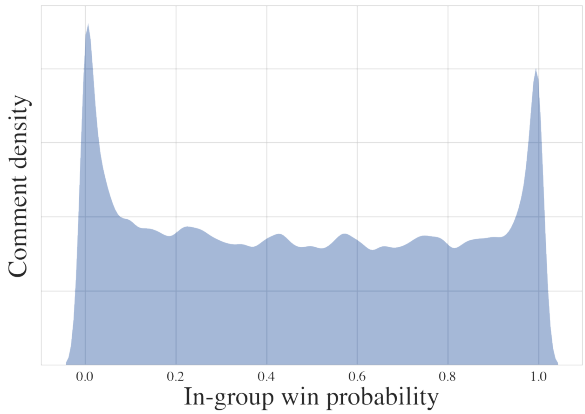

Figure 5: Comment density against WP.

B Annotation

Protocol Annotators were given the following instructions:

1. All comments are from game threads corresponding to specific NFL games between two teams. You will be given the source of the comment — this is the team the writer of the comment supports, the opponent in that game, and the live score at the time of making the comment.
2. Highlight any words and phrases that refer to individuals (people, teams, sub-groups within the team, organizations).
3. If the reference is to the same group as the source subreddit of the comment, tag this highlight as **in-group** ([IN]).
4. If the reference is towards the opponent in this specific game for which the comment is written, tag this highlight as **out-group** ([OUT]).
5. If the reference is towards any other team in the NFL apart from the two teams involved in this game, tag this highlight as **other** ([OTHER]).
6. Some comments will not have an obvious reference to an in-group/out-group/other entity. Leave these comments un-annotated. If you're unsure of an annotation, you can indicate your confidence, but only use the confidence scale if you're not very confident with your annotation. I will take an empty confidence annotation as full confidence.
7. Do not annotate a [SENT] token if there is a word in the sentence that can be annotated with the same label.

They were also given the following examples. Models were finetuned with the following as few-shot examples — they were provided WP over live score for a more holistic representation of the game, and explanations were modified depending on whether WP was provided or not.

Example 1

```
COMMENT: [SENT] Defense getting
absolutely bullied by a dude that
looks like he sells solar panels
IN-GROUP: Jets
OUT-GROUP: Bears
LIVE SCORE: Jets 7 - 3 Bears
TARGET: [SENT] [IN] getting
absolutely bullied by [OUT] that
looks like [OUT] sells solar panels.
EXPLANATION: The commenter is
probably talking about the in-group,
since 'Defense' is said without
qualification, and the description of
the offensive player is disparaging
('he sells solar panels'). 'Defense'
should be tagged [IN] since it refers
to in-group, and 'a dude' and 'he'
should be tagged [OUT] since it refers
to an out-group offensive player.
```

Example 2

```
COMMENT: [SENT] Hasn't really been him .
[SENT] Receivers have been missing a lot
of easy catches.
IN-GROUP: Dolphins
OUT-GROUP: Chargers
LIVE SCORE: Dolphins 0 - Chargers 0
TARGET: [SENT] Hasn't really been [IN] .
[SENT] [IN] have been missing a lot of
easy catches .
EXPLANATION: The second sentence is
complaining about the receivers missing
a lot of catches, thus absolving another
```

player of some blame, which is something fans would only do for the in-group team they support. Thus 'him' in first sentence, and 'Receivers' in second sentence should be tagged with [IN].

Example 3

COMMENT: [SENT] Cards and rams are gonna be in the post-season regardless, so I don't really care about them losing unless they play us.
IN-GROUP: 49ers
OUT-GROUP: Jaguars
LIVE SCORE: 49ers 30 - 10 Jaguars
TARGET: [SENT] [OTHER] and [OTHER] are gonna be in the post-season regardless, so I don't really care about [OTHER] losing unless they play [IN].
EXPLANATION: The game is between the 49ers and Jaguars, while the words 'Cards' and 'rams' refers to other teams in the NFL. Thus they should be tagged [OTHER] since they are neither in-group nor out-group, as should the word 'them'. 'us' should be tagged [IN] since it refers to the in-group team the player supports.

Example 4

COMMENT: [SENT] How are we this shit on defense
IN-GROUP: Steelers
OUT-GROUP: Eagles
LIVE SCORE: Steelers 7 - 21 Eagles
TARGET: [SENT] How are [IN] this shit on defense
EXPLANATION: 'we' here, and almost always, refers to the in-group since they don't like their team's defense, which is reflected in the score. 'we' should therefore be tagged with [IN] since it refers to in-group.

Example 5

COMMENT: [SENT] The chiefs got straight fucked with that Herbert INT getting called dead .
[SENT] Suck it , KC !
IN-GROUP: Chargers
OUT-GROUP: Chiefs
LIVE SCORE: Chargers 28 - 28 Chiefs
TARGET: [SENT] [OUT] got straight fucked with that [IN] INT getting called dead . [SENT] Suck it , [OUT] !
EXPLANATION: This is a game between the Chiefs and the Chargers, and the commenter is a supporter of the Chiefs, so 'the chiefs' in the first sentence and 'KC' in the second sentence should be tagged [OUT]. Herbert is a player for the Chargers, and should be tagged with [IN] since he is a member of the in-group with respect to the commenter.

Example 6

COMMENT: [SENT] Need points but 7 would be HUGE momentum
IN-GROUP: Bengals
OUT-GROUP: Chiefs
LIVE SCORE: Bengals 3 - 13 Chiefs
TARGET: [IN] Need points but 7 would be HUGE momentum
EXPLANATION: The in-group team is losing currently as the score shows, so this comment is implicitly about the in-group needing points to gain momentum. Thus '[SENT]' should be tagged with '[IN]' since there is no explicit word/phrase that refers to the in-group, but the comment is referring to the in-group implicitly.

C Prompts

Below is the prompt provided to both GPT-4o and Llama-3. Examples are the same as the ones provided to human annotators, listed in the previous section. The following prompt does not use win probabilities; The prompts which do use WP are the same as below, except they include a definition of WP as 'the probability of the in-group winning the game at the time of the comment - if the win probability is high, the in-group team is probably doing well and going to win.' in the prompt text.

Tag references to entities as in-group ([IN]), out-group ([OUT]) or other ([OTHER]) in live, online sports comments during NFL games. The input is the comment, the in-group team the commenter supports and the out-group opponent team during that game. Using knowledge of American football and contextual language understanding, identify words and phrases denoting entities (players, teams, city names, sub-groups within the team) that

```
refer to the in-group ([IN] - team the
commenter supports), out-group ([OUT] -
the opponent) or other teams ([OTHER] -
some other team in the NFL that is not the
in-group or the opponent), with respect
to the commenter. Return the list of
words/phrases that are to be tagged
(REF_EXPRESSIONS), an EXPLANATION
reasoning over why these words and phrases
in COMMENT should be tagged and with what
tag, and the TARGET comment itself with
relevant words/phrases replaced with the
respective tags ([IN], [OUT] or [OTHER])
in your final output.

Each sentence in a comment is separated by
a [SENT] token. Sometimes a sentence in
the comment will be about the in/out/other
group but not have an explicit word/phrase
that refers to the group; In such cases,
tag the [SENT] token for that sentence with
the corresponding tag label.

Here are 6 examples, with
REF_EXPRESSIONS being the list of
words/phrases to be tagged from COMMENT,
EXPLANATION being a reasonable reason
for why these words/phrases should be
tagged with appropriate tags, and TARGET
being the correct tagged output for
COMMENT.

[EXAMPLES 1-6 FOLLOW HERE]

Some comments will have no explicit or
implicit reference to the in-group,
out-group, or other, or it could be
extremely hard to disambiguate any
references based on given information.
In such cases, return Target as a copy
of Comment, justify this with the
Explanation, "No explicit or implicit
references to tag.", and return [] for
REF_EXPRESSIONS. Here is an example:

COMMENT: [SENT] I thought so. [SENT]
Wish I could say the same ;)
IN-GROUP: Jaguars
OUT-GROUP: Titans
REF_EXPRESSIONS: []
EXPLANATION: No explicit or implicit
references to tag.
TARGET: [SENT] I thought so. [SENT]
Wish I could say the same ;)

Now tag only the relevant words/phrases
in the following comment as either
in-group ([IN]), out-group ([OUT]), or
other ([OTHER]), if any. First return
the list of words to be tagged, then
explain your reasoning as to why these
words/phrases should be tagged from
COMMENT and with which tags, and finally
return the tagged comment in that order.
```

The explanations in the prompt with WP are similar to the explanations provided previously. Here is the explanation for Example 1 from the previous section using win probabilities:

```
The commenter is probably talking about
the in-group, since 'Defense' is said
without qualification, and the
description of the offensive player is
disparaging ('he sells solar panels').
This is in spite of the win probability
being relatively high for the in-group
- sometimes commenters choose to focus
on immediate plays rather than the
overall state of the game, so perhaps
this comment was in response to a bad
showing by the defense. 'Defense' should
be tagged [IN] since it refers to
in-group, and 'a dude' and 'he' should
be tagged [OUT] since it refers to an
out-group offensive player.
```

Here is an example model input after the prompt:

```
COMMENT: [SENT] Defense getting
absolutely bullied by a dude that
looks like he sells solar panels.
IN-GROUP: Jets
OUT-GROUP: Bears
WIN PROBABILITY: 71.5%
REF_EXPRESSIONS:
```

and here is the model's expected output:

```
['Defense', 'a dude', 'he']
EXPLANATION: The commenter is
probably talking about the in-group,
since 'Defense' is said without
qualification, and the description
of the offensive player is disparaging
('he sells solar panels'). 'Defense'
should be tagged [IN] since it refers
```

to in-group, and 'a dude' and 'he' should be tagged [OUT] since it refers to an out-group offensive player. TARGET: [SENT] [IN] getting absolutely bullied by [OUT] that looks like [OUT] sells solar panels .

Explanations from GPT-4o were generated using the same prompt format — the model is tasked with generating EXPLANATION instead of TARGET. The same examples were used, with EXPLANATION at the end rather than TARGET. GPT-4o was tasked to generate explanations of maximum length 256 tokens with seed of 1.

Linguistic WP In the linguistic WP condition, we replace the percentage WP value with a text string like below:

- 0–25: *Team A* is very likely to lose.
- 25–45: *Team A* is likely to lose.
- 45–55: Both teams are equally likely to win.
- 55–75: *Team B* is likely to win.
- 75–100: *Team B* is very likely to win.

D Modeling implementation

GPT-4o All few-shot experiments were run with `gpt-4o-2024-05-13`. Temperature was set to 1 if temperature scaling wasn't used, else it is dynamically set to $sin(\pi \times WP)$.

Llama-3-8B We fine-tuned the base `llama-3-8b` model from Meta's Huggingface model space[2]. We used the `Axolotl`[3] framework for all fine-tuning experiments with the following hyper-parameter settings:

- batch size of 4 for training and inference.
- sample packing and padding to sequence length were enabled, with a max sequence length of 2560. None of our inputs exceeded this limit.
- Cosine learning rate scheduler with warmup of 10 steps, learning rate set to $1e-5$, weight decay of 0.1, and a minimum learning rate ratio of 0.1
- Maximum of 2 train epochs with early stopping, and patience set to 3.
- The model is evaluated and saved every 59 steps for a maximum of 595 steps.
- Flash attention and gradient checkpointing were enabled.

[2] huggingface.co/meta-llama/Meta-Llama-3-8B

[3] https://github.com/OpenAccess-AI-Collective/axolotl

All finetuning experiments were done on 2 Nvidia A40 GPUs, and each fine-tuning run took approximately 1.5 hours.

E Modeling Analysis

Feature	Slope($\times 10^{-4}$)	R-squared
Any reference	-19.3	0.72
No reference	2.4	0.65
In-group	-2.8	0.31
we	-2	0.61
out-group	2.5	0.56
they_in	-0.3	0.15
they_out	0.4	0.25

Table 2: Table of slopes of feature of interest against increasing WP, alongside the r-squared showing how much of the variance is explained by the linear regression fit. The slopes for Any and no reference are calculated with frequencies normalized by total number of referents in a WP window. All other slopes for referent variables are measured with frequencies normalized by comments with references in that WP window.