It's Not Bragging If You Can Back It Up: Can LLMs Understand Braggings?

Jingjie Zeng[1], Huayang Li[1], Liang Yang[1,2]✉*, Yuanyuan Sun[1], Hongfei Lin[1]
[1]School of Computer Science and Technology, Dalian University of Technology, China
[2]Key Laboratory of Social Computing and Cognitive Intelligence, Ministry of Education, China
jjtail@mail.dlut.edu.cn, ✉liang@dlut.edu.cn

Abstract

Bragging, as a pervasive social-linguistic phenomenon, reflects complex human interaction patterns. However, the understanding and generation of appropriate bragging behavior in large language models (LLMs) remains underexplored. In this paper, we propose a comprehensive study that combines analytical and controllable approaches to examine bragging in LLMs. We design three tasks, *bragging recognition, bragging explanation, and bragging generation*, along with novel evaluation metrics to assess the models' ability to identify bragging intent, social appropriateness, and account for context sensitivity. Our analysis reveals the challenges of bragging in the social context, such as recognizing bragging and responding appropriately with bragging in conversation. This work provides new insights into how LLMs process bragging and highlights the need for more research on generating contextually appropriate behavior in LLMs[1].

1 Introduction

Understanding subtle social behaviors, such as bragging, is essential for improving large language models (LLMs) in socially sensitive applications. Bragging is a complex social behavior in which individuals highlight their achievements or qualities, often with the intention of overstating their social status, and therefore it is frequently viewed negatively in certain contexts (Festinger, 1954; Goffman, 1959; Leary, 1995). There are various forms of bragging, ranging from overt self-promotion (Dayter, 2014) to more subtle forms like humblebragging (Sezer et al., 2018). It represents a deliberate strategy of impression management and self-presentation. The effectiveness and appropriateness of this behavior are deeply rooted in cultural norms (Tannen, 1990), social contexts (Cheng et al., 2010), and individual characteristics (Paulhus and John, 1998), making it a subtle aspect of human social interaction (Cheng et al., 2010).

Figure 1: Toy examples of response bragging, which illustrates how three different LLMs respond to a user's bragging statement about a promotion. $\mathbf{LLM}_A$ offers a supportive response, $\mathbf{LLM}_B$ provides a balanced response with a warning, and $\mathbf{LLM}_C$ mirrors the user's bragging. The users seeking validation through bragging might prefer $\mathbf{LLM}_C$'s approach, even though $\mathbf{LLM}_B$ offers more constructive feedback.

The social appropriateness of bragging is highly context-dependent (Fiske, 1993), which requires a sophisticated understanding of social norms (Benedict, 1946; Goffman, 1959), relationship dynamics (Leary, 1995; Cheng et al., 2010), and cultural expectations. A successful brag often involves careful calibration of the content, timing (Tannen, 1990; Leary and Kowalski, 1990), and delivery method (Duncan Jr, 1969)—factors that even humans sometimes struggle to navigate properly.

As LLMs increasingly engage in social interactions, their ability to understand and appropriately handle bragging behavior becomes crucial.

Consider the scenario as shown in Figure 1. Where a user bragging announces *"**I just got promoted - I deserve it more than anyone else in the office,**"* how should LLMs respond? This seemingly simple interaction encapsulates the complex challenges that language models face in navigat-

*Corresponding Author
[1]The Code and datasets are available at https://github.com/jjtail/Bragging_acl2025

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 17542–17560
July 27 - August 1, 2025

ing the social nuances of bragging. The different LLMs responses to the announcement: $\mathbf{LLM}_A$ offers a simple congratulatory response. While professionally appropriate, this response fails to engage with the underlying social dynamics of user's bragging behavior. $\mathbf{LLM}_B$ takes a teachable approach, the response of which demonstrates social awareness by acknowledging the achievement while gently addressing the potentially problematic comparative mindset. $\mathbf{LLM}_C$ responds playfully, while this response resonate with users seeking validation. This variability highlights the intricate balance LLMs must achieve to respond appropriately to bragging in diverse contexts.

Despite bragging is very common in human communication and the increasing deployment of LLMs in social contexts, current research on how LLMs process and generate bragging behavior remains limited. Existing studies, such as the work by Jin et al. (2022) have created datasets based on Twitter (now X) and largely defined bragging as a subtask under the broader category of social knowledge, particularly related to trustworthiness (Choi et al., 2023). However, these studies have focused mainly on bragging recognition classification tasks (Choi et al., 2023; Li and Zhou, 2023; Mu et al., 2024b,a), with little exploration on how LLMs understand bragging, let alone how they might generate bragging-related content.

This gap motivates the following **R**esearch **Q**uestions, aligned with the proposed tasks of *Bragging Recognition*, *Bragging Explanation* and *Bragging Generation*:

RQ 1: What is the performance trend across different LLMs on bragging recognition tasks?

RQ 2: What extent can LLMs accurately explain the key sociolinguistic components of bragging?

RQ 3: How do LLMs generate the bragging-related contents and how do their responses align with user expectations and social norms?

To address these questions, we conduct a series of experiments in the following part of this work. Section 4 addresses LLMs' challenges in bragging recognition, Section 5 explores their understanding of bragging, and Section 6 investigates the generation of bragging behavior, focusing on a two-stage evaluation that measures both the generation of contextually appropriate bragging and the responses to bragging in simulated conversations. In summary, our contributions are as follows:

• To the best of our knowledge, this is the first work to systematically evaluate LLMs' capabilities in bragging behavior. We are not just analyzing existing data, but also exploring how LLMs can understand and generate bragging, which is a novel and challenging research area.

• We propose new evaluation metrics to quantify bragging intensity, contextual appropriateness, and social impact that haven't been measured before in the context of LLMs.

• Through extensive experiments across various LLMs, we provide a thorough analysis of the results. Our findings reveal the limitations of LLMs in handling complex social behaviors like bragging, especially in adapting to different contexts and interpreting implicit bragging.

2 Related Work

Bragging, recognized as a significant linguistic art of language, has attracted considerable attention in academic research across diverse fields such as linguistics, psychology and sociology (Leech, 2014; Scopelliti et al., 2015; Sezer et al., 2018; Chaudhry, 2019). These studies analyze bragging by examining its illocutionary force, its underlying structure, and its intricate relationship to other language acts. Significantly, humble-bragging is increasingly perceived as a sophisticated trick and a potentially advantageous tool for self-promotion (Harrison, 2024; Matley, 2018).

In the realm of AI, with the rise of LLMs, interactions between humans and LLMs are gradually increasing. This shift needs a new dimension to our understanding of bragging. Jin et al. (2022) create a bragging dataset based on Twitter (now X), which includes bragging statements and various classification tasks related to bragging types. Based on this, Choi et al. (2023) investigate the performance of LLMs on several social knowledge datasets, incorporating bragging as one of the subtasks under trustworthiness. Mu et al. (2024a,b) conduct a study on the zero-shot classification accuracy of LLMs for the bragging task, finding that LLMs perform worse than some pre-trained models in classifying bragging statements. From here we see that, current research about bragging with LLMs has not addressed the fundamental question: How should LLMs respond to bragging?

A key aspect of appropriate social interaction involves responding to others' self-disclosures, including bragging, in a balanced manner. Therefore, in this paper, when evaluating LLM responses to bragging, we also assess whether they avoid unde-

sirable extremes such as sycophancy or excessive preachiness, aiming for responses that are socially congruent rather than simply agreeable. Future work could explore how to actively guide LLMs towards such balanced responses.

3 Preliminaries and Tasks Definition

In this section, we design an evaluation method consisting of three progressive tasks to evaluate whether LLMs can understand bragging well.

3.1 Task Formulation

Task 1: Bragging Recognition This task assesses an LLM's ability to identify bragging statements within a given text. It involves classifying text as either "bragging" or "non-bragging", as shown in the following two examples.

> **Input Text**: My schedule is so packed, it is hard to make time for award dinners.
> **Model Output**: The given text is a bragging.
> **Input Text**: I have a great workout today.
> **Model Output**: The given text is a non-bragging.

Task 2: Bragging Explanation This task assesses an LLM's understanding of the social context and motivations behind bragging statements. It involves generating explanations for why a given statement is considered bragging and evaluating its social appropriateness. Here is an example:

> **Input Text**: I'm so exhausted from all these award ceremonies I have to attend.
> **Model Output**: The speaker is pretending to complain about being tired but is actually drawing attention to the fact that they are attending many award ceremonies, implying they are successful and in-demand. This is a common and relatively acceptable form of bragging in this context.

Task 3: Bragging Generation This task assesses an LLM's ability to generate bragging-related content, either in response to a prompt or as part of a dialogue. We focus on two distinct scenarios: ***(i)*** generating bragging content based on a specific prompt and social context,

> **Description**: In this scenario, the LLM receives a prompt that instructs it to generate a bragging statement related to a specific topic or situation. Crucially, the model must tailor its response to a specified social context.
> **Input Prompt**: letting interviewer remember.
> **Social Contextual**: job seekers
> **Model Output**: I have a proven track record of turning caffeine input into productivity output.

and ***(ii)*** responding appropriately to a user's bragging statement.

> **Description**: This scenario focuses on the LLM's ability to respond appropriately when a user makes a bragging statement. The user's statement will always be a form of bragging. Unlike Scenario 1, social context is not explicitly provided in this scenario, as the focus is on general appropriateness in responding to bragging.
> **Input Prompt**: Respond to: "I just closed the biggest deal in the company's history!"
> **Model Output**: Wow, that's a major accomplishment! You must have put in a ton of work.

3.2 Task Implementation

We design specific prompts for LLMs to test their inherent abilities on these three tasks[2].

• For **bragging recognition**, the focus is on assessing the model's accuracy and confidence in classifying statements as bragging or non-bragging. To test the model's robustness, we include two slightly biased instructions, one favoring bragging and the other favoring non-bragging. This approach evaluates the model's ability to handle subtle biases and varying contextual cues effectively.

• For **bragging explanation**, we employ the Chain-of-Thought (CoT) technique (Wei et al., 2022), which prompts the model to provide a reason before classifying statements. The generated reason is directly used to evaluate the model's ability to explain bragging behavior. By integrating CoT, we ensure the model engages in multi-step reasoning, which enhances the quality of explanations.

• For **bragging generation**, we evaluate the model's ability to create bragging-related content in two scenarios. *1)* ***Prompt-driven Bragging Generation***, where the model is tasked with crafting a bragging statement based on a specific prompt and social context; and *2)* ***Responding to User Bragging***, where the model is required to appropriately respond to a user's bragging statement. This scenario tests the model's ability to balance politeness and engagement.

3.3 Data Selection and Construction

We utilize the publicly available bragging dataset provided by Jin et al. (2022), which is originally constructed from tagged Twitter data (*e.g., #humblebrag, #bragging*) for classification tasks. However, as this dataset is not designed to include explanations or information on the social and contextual aspects of bragging, we re-annotate the bragging samples to incorporate such details[3]. For our tasks,

[2] All prompts for three bragging-related tasks are available at Appendix A.

[3] The specifics of the re-annotation process, including the explanation and social context, are provided in Appendix B.

Task	Examples (Bragging / non-Bragging)	Test Data (Bragging / non-Bragging)
Recognition	10 / 10	781 / 5915
Explanation	3 / 0	781 / 0
Generation	3 / 0	200 / 0

Table 1: The data splits for the bragging tasks are as follows: **Recognition** uses the original dataset from Jin et al. (2022), **Explanation** is a re-annotated subset with added explanations and social context, and **Generation** uses a subset of the re-annotated data from the Explanation task, focusing on Scenario 1: *Prompt-driven Bragging Generation*.

we apply different segmented dataset based on the specific requirements of each task and select a subset of examples as demonstrations in the prompts, as shown in Table 1.

3.4 Model Selection

To evaluate the level of understanding of bragging across LLMs with varying parameter sizes and capabilities, we selected eight well-known LLMs for experimentation, divided into two categories. The first category includes open-weight models with parameter sizes under 10B, such as Llama3.1-8B-Instruct (Meta, 2024), Gemma-2-9B-It (Gemma, 2024), Mistral-8B-Instruct (Mistral, 2024), and Qwen2.5-7B-Instruct (Qwen, 2024). The second category consists of closed-source models with larger parameter scales, such as Gemini-2.0-Flash (DeepMind, 2024), ChatGPT-4o-Latest (OpenAI, 2024a), Claude-3.5-Sonnet-20241022 (Anthropic, 2024), and o1-mini-2024-09-12 (OpenAI, 2024b). All these models are generative text models equipped with context learning and instruction-following capabilities.

3.5 Evaluation Metrics

Metrics for Recognition To evaluate the performance of LLMs on the Bragging Recognition task, we utilize four key metrics: 1) **True Positive Rate (TPR)**, which measures the proportion of correctly identified bragging statements; 2) **True Negative Rate (TNR)**, which measures the proporction of correctly identified non-bragging statements; and 3) **Accuracy (Acc)**, which measures the overall correctness of the classifications. We calculate Acc as the average accuracy across the two biased prompts (one leaning towards bragging and the other towards non-bragging). Additionally, we calculate the changes Δ**TPR** and Δ**TNR**, which quantify the model's consistency and susceptibility to prompt bias by measuring the difference in TPR and TNR when the prompt's leaning is shifted from bragging to non-bragging.

Metrics for Explanation Assessing the quality of bragging explanations is inherently complex due to the subtle nature of social interactions and the subjective interpretation of language. Therefore, we employ a multi-faceted evaluation approach that combines fine-grained human assessment with large-scale automatic evaluation using pairwise comparisons. *1)* ***Fine-grained Element Identification Check***. We randomly sample 100 bragging statements from our dataset. For each statement, we present the statement and the LLM-generated explanation to three human annotators. The annotators are instructed to evaluate whether the explanation correctly identifies and mentions the following key elements of bragging: **Potential Social Context** (Hymes, 1974; Goffman, 1959), **Speaker's Intention** (Searle et al., 1980), **Desired Feedback** (Jones, 1982), and **Appropriateness** (Brown, 1987). For each of the four elements[4], we calculate the percentage of explanations where each annotator agrees that the element is correctly identified. We then compute the average agreement across all elements. *2)* ***Large-scale Pairwise Comparison***. We use the human-annotated explanations from our re-annotated dataset (described in Section 3.3) as the gold standard. For each bragging statement in the dataset, we present GPT-4 with two explanations: one generated by an LLM and the other written by a human annotator. GPT-4 is instructed to act as an impartial judge and select the better explanation. We calculate the **Win Rate**, **Tie Rate**, and **Loss Rate**. This approach is widely used for evaluation (Qin et al., 2024; Xu et al., 2024). However, we acknowledge the inherent limitations of using the LLM as a definitive "impartial judge," as discussed further in Section 8.

Metrics for Generation Our evaluation of the Bragging Generation task focuses on two distinct scenarios, recognizing the critical role of context and interaction in assessing bragging behavior. Scenario 1 serves as a prerequisite for Scenario 2, allowing us to evaluate the generated bragging statements both in isolation and in the context of a subsequent interaction.

(i) Prompt-driven Bragging Generation. Requiring LLMs to produce bragging statements given

[4]We provide a detail for the selection of these metrics for bragging explanation in Appendix C.

specific prompts and social contexts. We evaluate the generated content using both human and automatic evaluation methods. Human annotators assess four key dimensions—**Bragging Success** (Sezer et al., 2018), **Complied Social Context** (Goffman, 1959), **Social Goal AchievMement** (Jones, 1982), and **Bragging Intensity**. While automatic metrics measure **huMor** (Baranov et al., 2023), a common element in successful bragging. We then use these generated statements as inputs for Scenario 2, allowing us to indirectly evaluate their quality based on how the LLMs responds to them as a listener.

(ii) Responding to User Bragging. Scenario 2 leverages the bragging statements generated by the LLMs in Scenario 1 as input. Here, the LLM is prompted to engage with these previously generated bragging statements, enabling a conversational setting. We utilize a sentiment analysis model (Sanh, 2019) (fine-tuned on SST-2 (Socher et al., 2013)) to measure how well the LLM's responses emotionally align with the user's bragging—ideally reflecting acknowledgment or validation, named **Sentiment Gap** (Hatfield et al., 1993). Human annotators also assess **Preachiness Intensity (PI)** (Brown, 1987) and **Sycophancy Intensity (SI)** (Vonk, 2002) in the responses[5].

In this way, Scenario 2 serves as a practical "stress test" for the outputs of Scenario 1, allowing us to link the quality of generated bragging statements directly to how the same LLM interacts with them in a conversational setting. By combining both human and automatic evaluations across these two scenarios, we aim to comprehensively assess the LLM's capacity to generate and respond to bragging in a manner that is contextually appropriate and socially friendly.

4 Bragging Recognition

This section details the experimental setup and results for the Bragging Recognition task. We evaluate eight LLMs on their ability to identify bragging statements, focusing on how prompt bias affects their performance when using a CoT approach.

Prompt Setting Each prompt provides a definition of bragging and required the models to classify input text as either "bragging" or "non-bragging". The provided reason follows a final classification in a structured JSON format[6]. The prompts are intentionally designed to bias the models towards a specific classification: one skewed towards bragging and the other towards non-bragging. The bragging biased prompt includes only examples of bragging statements, while the non-bragging biased prompt provides only non-bragging examples. This design allows us to analyze the models' contextual bias within the prompts and its impact on their ability to accurately recognize bragging.

Model	Bragging -> Non-Bragging				
	TPR	Δ TPR	TNR	Δ TNR	Acc
Llama3.1-8B-It	**0.826**	-0.264	0.380	+0.365	0.578
Gemma-2-9B-It	0.576	-0.167	0.783	+0.065	0.778
Mistral-8B-It	0.696	-0.273	0.669	+0.201	0.744
Qwen2.5-7B-It	0.551	-0.244	0.857	+0.064	0.835
Gemini-2.0-Flash	0.571	**-0.020**	0.829	+0.078	0.836
ChatGPT-4o-Latest	0.541	-0.137	0.910	+0.056	**0.890**
Claude-3.5-Sonnet	0.242	+0.300	**0.972**	-0.062	0.882
o1-mini	0.523	-0.042	0.910	**+0.010**	0.863

Table 2: Results of two biased bragging recognition. In addition to calculating TPR, TNR, and Acc, we also measure the changes Δ in TPR and TNR when the prompt bias transitions from bragging to non-bragging. Acc as the average accuracy across the two biased prompts. The best results are **bolded**, while the second results are underlined.

Recognition Result Analysis Table 2 presents the each model under the two prompt conditions. We can find that:

*1) **Influence of Biased Prompts**:* The results clearly show that biased prompts significantly affect model performance. As expected, transitioning from bragging to non-bragging-biased prompts decreases TPR and increased TNR across most of the models. This sensitivity isn't merely an experimental artifact; it reflects real-world LLM usage where examples often align with desired outcomes. This highlights the potential pitfalls of such common practices.

*2) **Unique Performance of Claude-3.5-Sonnet**:* The model's performance on the bragging recognition task is notably distinctive. When prompted with bragging cues, it shows a low TPR of 0.242 and a high TNR of 0.972, indicating a strong bias toward classifying statements as non-bragging. In contrast, when prompted with non-bragging cues, the TPR rises to 0.542, while the TNR slightly drops to 0.910. This behavior likely stems from Claude-3.5-Sonnet's initial training on synthetic data with a narrow definition of bragging, resulting

[5]We provide a detail for the selection of these metrics for bragging generation in Appendix D.

[6]The details of prompts are shown in Appendix A.

in a high threshold for classification. The non-bragging examples may have helped refine this threshold, improving its ability to recognize subtler forms of bragging.

3) ***Poor Performance and Lack of Robustness****:* The observed performance across all models is relatively poor, further emphasized by the consistently low accuracy values. This suggests a general lack of robustness and a high sensitivity to prompt bias. These findings align with observations from previous work (Mu et al., 2024a,b), which finds that pre-trained models specifically fine-tuned for tasks like bragging detection outperform LLMs.

4) ***The challenge of Bragging Recognition****:* Bragging on social media is complex, often expressed through informal language including abbreviations, slang, and emojis. Such data is inherently noisy and ambiguous, creating a significant obstacle for LLMs. Bragging is a subtle social behavior influenced by various underlying motivations and social goals. It is frequently characterized by a level of intentional ambiguity, exemplified by phenomena like humble-bragging or indirect self-promotion. The subpar performance, even with CoT prompting, suggests that current LLMs lack the necessary in-depth understanding to reliably identify bragging. They appear influenced by surface-level cues within the prompt, rather than exhibiting a genuine grasp of the underlying social interactions.

Connecting Results to RQ 1: **"What is the performance trend across different LLMs on bragging recognition tasks?"** While some prior studies (Choi et al., 2023; Li and Zhou, 2023; Mu et al., 2024b,a) have touched upon LLMs' performance in bragging classification, this research question remains crucial. Firstly, our study examines a broader and more current range of LLMs, including recent open-weight and advanced closed-source models not extensively covered previously. Secondly, and more importantly, RQ1 specifically investigates the robustness of these models to prompt bias—a critical aspect of real-world LLM interaction that can reveal deeper limitations than simple classification accuracy. The findings from this investigation into prompt sensitivity, which serve as a vital baseline for understanding the more complex tasks of bragging explanation and generation explored in subsequent sections, reveal that LLMs generally struggle with recognizing bragging. They exhibit significant prompt bias and low accuracy, suggesting a reliance on superficial patterns rather than a genuine understanding of the underlying social interactions inherent in bragging. This vulnerability to contextual manipulations within prompts underscores their current inability to engage in deeper social reasoning, highlighting the challenges in creating LLMs capable of reliably understanding subtle social cues like those present in bragging, a key implication for real-world applications.

Model	Fine-grained Check			
	Context	Intention	Feedback	APP
Llama3.1-8B-It	0.47	0.84	0.73	0.44
Gemma-2-9B-It	0.46	0.88	0.68	0.51
Mistral-8B-It	0.40	0.74	0.61	0.38
Qwen2.5-7B-It	0.48	0.86	0.69	0.43
Gemini-2.0-Flash	<u>0.57</u>	<u>0.94</u>	<u>0.85</u>	<u>0.55</u>
ChatGPT-4o-Latest	0.52	0.88	0.74	0.47
Claude-3.5-Sonnet	0.51	0.92	0.72	0.42
o1-mini	0.46	0.89	0.73	0.48
Human	**0.92**	**0.98**	**0.93**	**0.85**

Table 3: Results of the *Fine-grained Element Identification Check*, evaluated by humans for the bragging explanation, include Potential Social **Context**, Speaker's **Intention**, Desired **Feedback**, and **App**ropriateness. The best results are **bolded**, while the second-best results are <u>underlined</u>.

5 Bragging Explanation

This section evaluates the ability of LLMs to explain why a given statement is considered bragging, focusing on their understanding of the social context and motivations behind such statements. We employ both fine-grained element identification checks and large-scale pairwise comparisons to assess the quality of the generated explanations.

Fine-grained Element Identification Check We present the results of the fine-grained element identification check in Table 3. This evaluation is based on 100 randomly sampled bragging statements and assesses whether the LLM-generated explanations correctly identify four key elements.

Large-scale Pairwise Comparison Unlike the fine-grained element identification, pairwise comparison focuses on the overall quality of explanations. We use human-annotated explanations from our re-annotated dataset as the gold standard. For each bragging statement, GPT-4 is tasked with choosing the better explanation between an LLM-generated one and a human-written one.

Explanation Result Analysis Combining Table 3 and Figure 2, we can draw the following

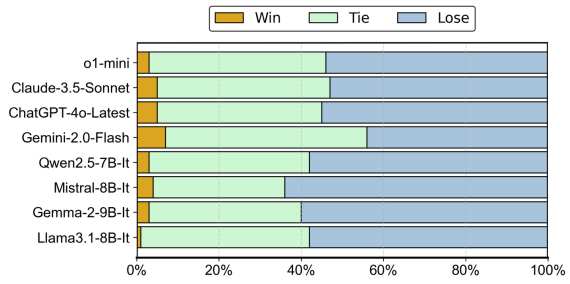

Figure 2: Evaluation of LLM-generated bragging explanations through large-scale pairwise comparison against human-written explanations. This figure presents the win, tie, and loss rates for eight evaluated LLMs. Each bar group represents an LLM, and the segments show the proportion of instances where the LLM's explanation was preferred over (Win), considered equal to (Tie), or deemed inferior to (Loss) the human gold standard by a GPT-4 judge, reflecting the models' ability to articulate the nuances of bragging.

conclusions:

1) ***Performance Gap*** There's a significant performance gap between all LLMs and human performance across all four elements. Humans demonstrate a much stronger understanding of the nuances of bragging explanations.

2) ***Partial Understanding of Bragging*** LLMs perform relatively well in identifying the Speaker's Intention (with most scores above 0.8). This suggests they can often grasp the underlying motivation behind a bragging statement, which may involve seeking validation, impressing others, etc. Gemini-2.0-Flash is the best among LLMs. LLMs struggle more with identifying the Potential Social Context and Appropriateness of bragging statements, with most scores falling below 0.5. This indicates difficulties in understanding the situational factors that influence how bragging is perceived. The performance on Desired Feedback is mixed. While most models score between 0.6 and 0.75.

3) ***Internal State at the Expense of External Factors:*** LLMs exhibit a tendency to prioritize understanding the internal motivations behind a speaker's bragging, often correctly identifying their intention. However, they struggle to grasp the external social cues that determine whether bragging is appropriate, demonstrating a lack of situational awareness. This imbalance, with a focus on the speaker's internal state over the external social context, leads to difficulty in recognizing how setting, audience, and other factors influence the acceptability and effectiveness of bragging. Consequently, LLMs also show inconsistent performance in predicting the likely reception of bragging statements, revealing a limited understanding of how intentions translate into social outcomes.

4) ***Confirmation of Weakness****:* Connecting Bragging Explanation Deficiencies to Prompt Influence in Bragging Recognition In essence, the models are not truly "recognizing" bragging in the same way humans do. Instead, they are performing a form of sophisticated pattern matching, guided by the cues provided in the prompt. When those cues are modified, their performance shifts dramatically.

Connecting Results to RQ2: **"What extent can LLMs accurately explain the key sociolinguistic components of bragging?"** The findings suggest that LLMs have a limited and often superficial comprehension of bragging. They can, to some extent, identify the explicit intention behind a bragging statement—recognizing that someone is trying to boast about an achievement or quality. This is evidenced by the higher scores for Speaker's Intention in Table 3. However, their understanding appears to be incomplete, particularly when faced with the complex content. They seem to treat bragging as an isolated act rather than a social behavior.

6 Bragging Generation

This section investigates the ability of LLMs to generate bragging-related content, focusing on two distinct scenarios: ***(i.)*** *Prompt-driven Bragging Generation*, where models generate bragging statements based on specific prompts and social contexts, and ***(ii.)*** *Responding to User Bragging*, where models respond appropriately to user-generated bragging statements. These scenarios evaluate the models' capacity to produce contextually relevant bragging and to engage in a socially appropriate manner when faced with bragging from a user.

Prompt-driven Bragging Generation LLMs are tasked with generating bragging statements tailored to specific social contexts using annotations from our re-annotated dataset (Section 3.3). The dataset includes contexts like "job interview," "meeting with friends," and "online forum." These are paired with prompts based on the "Speaker's Intention" field, guiding the model to brag about a specific topic within the given context. This setup allows us to assess how well the models adapt their bragging to different social settings. For example, a prompt to "highlight academic achievements"

Model	Prompt-driven Bragging Generation					Responding to User Bragging		
	Success	Context	Achievement	Intensity	huMor	PI	SI	SenGap
Llama3.1-8B-It	7.36	8.09	6.82	5.45	0.59	1.45	2.73	0.47
Gemma-2-9B-It	7.35	8.00	6.85	6.05	**0.73**	**0.74**	4.26	0.49
Mistral-8B-It	7.38	7.94	6.88	5.88	0.50	1.25	2.40	0.26
Qwen2.5-7B-It	7.61	**8.17**	7.00	5.94	0.62	2.15	3.05	0.38
Gemini-2.0-Flash	7.44	7.44	7.06	6.28	0.62	1.87	3.53	0.66
ChatGPT-4o-Latest	7.65	7.90	**7.20**	5.90	0.48	1.21	4.11	0.31
Claude-3.5-Sonnet	7.50	7.11	6.61	**6.94**	0.60	3.11	**2.00**	**0.21**
o1-mini	**7.80**	7.60	7.10	6.20	0.46	1.11	3.50	0.33

Table 4: Performance of Large Language Models on Bragging Generation tasks across two scenarios: (i) *Prompt-driven Bragging Generation* and (ii) *Responding to User Bragging*. For Scenario 1, models were evaluated on Bragging Success (Success), Compliance with Social Context (Context), Social Goal Achievement (Achievement), Bragging Intensity (Intensity), and Humor (huMor). For Scenario 2, responses were assessed for Preachiness Intensity (PI), Sycophancy Intensity (SI), and Sentiment Gap (SenGap). The best results are **bolded**, and second-best results are underlined.

could be paired with both "meeting with friends" and "job interview," requiring the model to generate contextually bragging for each scenario.

Responding to User Bragging We utilize the bragging statements generated by the LLMs in above scenario 1 as input for this scenario. Each model is then prompted to respond to these statements as if it is a participant in a conversation. This evaluates its capacity to engage in a conversation where bragging is present, demonstrating social ability by providing appropriate responses.

Generation Result Analysis In table 4, we can draw the following conclusions:

1) No Single Model Dominates The varying performance of different models in assessing or generating bragging behavior stems from the complexity of bragging itself, which is highly context-dependent and influenced by factors like intention, expression. Current LLMs, despite diverse architectures and training data, may not fully grasp these subtle differences due to limitations in their social common sense and a potential lack of adequate representation of complex social interactions in their training data.

2) Low Intensity with Safety and Ethical Considerations: Most maintain an average level of bragging. LLMs tend to reduce their bragging intensity due to safety and ethical concerns. Unchecked bragging can come across as offensive, arrogant, or insensitive, damaging relationships and impacting public opinions. By limiting bragging, which help prevent harmful or socially inappropriate content.

3) Why Bragging Generation Might Seem Easier than Recognition/Explanation: LLMs might find it easier to mimic the surface features of bragging than to truly understand its complex social dynamics. They can generate text that resembles bragging based on patterns learned from their training data, even without grasping the nuances of when it's appropriate. This is because generating a bragging statement is a less constrained task than recognizing or explaining it, and the "success" metric might be satisfied superficially with just a positive mention of achievement.

4) Negative Correlation between PI and SI: The observed negative correlation between preachiness and sycophancy in LLMs' responses to bragging highlights a crucial trade-off in their social intelligence. Models like Claude, which avoid flattery , tend to sound preachy, while models like Gemma, which avoid being preachy, often become overly flattering. This suggests that LLMs struggle to find the subtle "sweet spot" in responding to bragging, which we need ensure that the model doesn't overcompensate by exhibiting another undesirable behavior.

Illustrative Case Study To provide a more qualitative understanding of the differences in bragging generation capabilities across models, we present a few illustrative examples.

> **Given the input prompt with a specified:**
> • ***Potential Social Context***: An online social media setting where users discuss a popular public figure or event.
> • ***Speaker's Intention***: To self-promote success in acquiring scarce tickets, with a degree of self-awareness regarding the boastful tone.
>
> **Output:**
> • ***Claude-3.5-Sonnet***: Ok fine, I'll be that person - just casually secured impossible-to-get floor seats for BOTH sold-out shows while everyone was panic-posting about

the queue. #sorrynotsorry
• ***Gemini-2.0-Flash***: Somehow, I ended up with meet-and-greet tickets AND front-row seats... I'm not even gonna pretend this isn't a flex. #BlessedFan
• ***ChatGPT-4o-Latest***: Finally got my hands on those exclusive tickets everyone's been talking about - not trying to brag (well, maybe just a little)!
• ***o1-mini***: Just managed to snag the last [Celebrity] concert tickets! Not trying to brag, but I'm beyond excited!
• ***Llama3.1-8B-It***: Just scored the last two tickets to the hottest event of the summer - I'm pretty sure it's just beginner's luck, but I'll take it.
• ***Gemma-2-9B-It***: Okay, I'm officially bragging, but I got tickets to [event] when they sold out in seconds. #FanLife-Goals #WorthTheHype
• ***Mistral-8B-It***: Got my hands on those elusive tickets! I'm basically a rockstar now, but hey, it's all about the fans, right?
• ***Qwen2.5-7B-It***: Finally got my hands on those elusive VIP concert tickets - can't believe my luck, really!

These examples showcase a spectrum of bragging styles. For instance, Claude adopts a more assertive and slightly confrontational humble-brag, while Gemini directly labels its statement as a "flex." Models like ChatGPT and o1-mini explicitly mention "not trying to brag" before proceeding to do so, indicating an attempt to mitigate the boast. In contrast, Llama employs a more classic humble-brag by attributing success to "beginner's luck." The inclusion of hashtags also varies, with some models using them to amplify the boastful sentiment or acknowledge the context (e.g., #sorrynotsorry).

Connecting Results to RQ3: **"How do LLMs generate bragging-related content, and how do their responses align with user expectations and social norms?"**

(i) Prompt-driven Bragging Generation Content: LLMs generally find it easier to generate bragging content than to recognize or interpret it. When asked to generate bragging statements, they can draw on their internal understanding of social cues and intentions to create exaggerated claims. However, their output appears somewhat restrained, as indicated by moderate polarity levels, suggesting they understand the form of bragging but are not fully lifted the constraints on them.

(ii) Responding to Bragging: When responding to user-generated bragging, most LLMs tend to align with the emotional tone of the bragger. Across different model families—such as those from OpenAI and Google—this tendency is consistent, though Google's models often exhibit a more enthusiastic tone. Additionally, there is a notable inverse relationship between preachiness and flattery. For instance, Claude's outputs show the more attempts to "educate" the user, the less it flatters them. These patterns align with observed users' experience experiences and highlight varying strategies LLMs employ in managing social expectations.

7 Conclusion

In this paper, we present a comprehensive study of bragging behavior in LLMs, exploring their capabilities in recognizing, explaining, and generating bragging content. Our analysis on three tasks and eight different LLMs reveals significant challenges and limitations in how these models handle the complexity of bragging in social interactions.

In the Bragging Recognition task, LLMs show sensitivity to prompt bias and a reliance on surface-level patterns, struggling to recognize bragging accurately. The Bragging Explanation task reveals further challenges in understanding context and social appropriateness. In the Bragging Generation task, LLMs produce restrained bragging, likely reflecting safety concerns, and their responses to bragging statements often oscillated between preachiness and sycophancy, struggling to strike a socially appropriate balance.

Our findings underscore the fact that bragging, as a social-linguistic phenomenon, is deeply intertwined with context, intention, and social norms, presenting a substantial challenge for LLMs. While these models can simulate aspects of bragging behavior, they currently lack the sophisticated social intelligence needed to fully comprehend and appropriately respond to it.

8 Limitations

This paper has explored LLMs' understanding of bragging, a complex social task even for humans, requiring them to infer the speaker's intent and respond appropriately. Several limitations should be acknowledged.

The use of GPT-4 for large-scale pairwise comparison in the Bragging Explanation task (Section 5) introduces its own set of limitations. While GPT-4 demonstrates strong language understanding, it is not a perfectly objective or "impartial judge"(Zheng et al., 2023; Kim et al., 2024). Its judgments may not fully capture the nuances a human evaluator specialized in sociolinguistics might discern. The results from this automated comparison should therefore be interpreted as indicative trends. In Section 6, we evaluated response "appropriateness" rather than defining a universal "best" response, as this is highly context dependent. Our metrics capture general social norms but may not

reflect context-specific ideals.

However, a more practical concern arises: When I am bragging with large language models in conversation to seek self validation, is it truly appropriate for models to assume an instructional stance?

9 Ethics Statement

Our research is embarking on a crucial exploration of how different large language models detect and respond to bragging in communication. This study represents a novel and ethically significant contribution to the field of language analysis. Our objective is not merely to assess the detection capabilities of these models but to push the boundaries of their understanding of self-promotion and boastful behavior. By generating bragging sentences using state-of-the-art models, we seek to examine their ability to accurately identify and respond to boastful expressions, evaluating the outcomes across multiple dimensions.

Given the sensitive nature of the subject, we are fully aware of the potential risks in terms of reinforcing negative stereotypes or promoting inappropriate behavior. To mitigate these concerns, we have implemented strict ethical protocols throughout our research process. All generated content will remain confined to controlled, academic environments, and no harmful content will be disseminated outside the scope of our study. Our goal is to ensure that this work serves as a constructive tool for enhancing communication technologies while safeguarding against the misuse of AI in perpetuating harmful behaviors.

10 Acknowledgments

This research is supported by the Natural Science Foundation of China (No. 62076046, 61702080, 62366040), the Key R&D Projects in Liaoning Province award numbers (2023JH26/10200015), the Fundamental Research Funds for the Central Universities (DUT24LAB123).

References

Anthropic. 2024. Claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet.

Alexander Baranov, Vladimir Kniazhevsky, and Pavel Braslavski. 2023. You told me that joke twice: A systematic investigation of transferability and robustness of humor detection models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13701–13715.

Ruth Benedict. 1946. The chrysanthemum and the sword: patterns of japanese culture.

Penelope Brown. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Shereen J Chaudhry. 2019. Thanking, apologizing, bragging, and blaming: Responsibility exchange theory and the currency of communication. *Psychological review*, 126(3):313.

Joey T Cheng, Jessica L Tracy, and Joseph Henrich. 2010. Pride, personality, and the evolutionary foundations of human social status. *Evolution and Human Behavior*, 31(5):334–347.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Daria Dayter. 2014. Self-praise in microblogging. *Journal of Pragmatics*, 61:91–102.

Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message.

Starkey Duncan Jr. 1969. Nonverbal communication. *Psychological bulletin*, 72(2):118.

Leon Festinger. 1954. A theory of social comparison processes. *Human relations*, 7(2):117–140.

ST Fiske. 1993. Controlling other people. the impact of power on stereotyping. *The American Psychologist*, 48(6):621–628.

Gemma. 2024. Gemma.

Erving Goffman. 1959. The presentation of self in everyday life.

Sara Harrison. 2024. A little humor and bragging could help you land your next job. Accessed: 2024-12-09.

Elaine Hatfield, John T Cacioppo, and Richard L Rapson. 1993. Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Dell Hymes. 1974. *Foundations in sociolinguistics: An ethnographic approach*. Routledge.

Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz, and Nikolaos Aletras. 2022. Automatic identification and classification of bragging in social media. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3945–3959, Dublin, Ireland. Association for Computational Linguistics.

EE Jones. 1982. Toward a general theory of strategic self presentation. *Psychological perspectives on the self/Erlbaum*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Mark R Leary. 1995. *Self-presentation: Impression management and interpersonal behavior*. Routledge.

Mark R Leary and Robin M Kowalski. 1990. Impression management: A literature review and two-component model. *Psychological bulletin*, 107(1):34.

Geoffrey Leech. 2014. *The pragmatics of politeness*. Oxford University Press.

Xiang Li and Yucheng Zhou. 2023. Disentangled and robust representation learning for bragging classification in social media. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

David Matley. 2018. "this is not a humblebrag, this is just a brag": The pragmatics of self-praise, hashtags and politeness in instagram posts. *Discourse, Context Media*, 22:30–38. Discourse of Social Tagging.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1.

Mistral. 2024. Introducing the world's best edge models. https://mistral.ai/news/ministraux/.

Yida Mu, Mali Jin, Xingyi Song, and Nikolaos Aletras. 2024a. Enhancing data quality through simple de-duplication: Navigating responsible computational social science research. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12477–12492, Miami, Florida, USA. Association for Computational Linguistics.

Yida Mu, Ben P. Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024b. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086, Torino, Italia. ELRA and ICCL.

OpenAI. 2024a. Hello gpt-4o. https://openai.com/index/hello-gpt-4o.

OpenAI. 2024b. Introducing openai o1-preview and o1-mini. https://help.openai.com/en/articles/9624314-model-release-notes.

Delroy L Paulhus and Oliver P John. 1998. Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of personality*, 66(6):1025–1060.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, dahai li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*.

Team Qwen. 2024. Qwen2.5: A party of foundation models.

V Sanh. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Irene Scopelliti, George Loewenstein, and Joachim Vosgerau. 2015. You call it "self-exuberance"; i call it "bragging" miscalibrated predictions of emotional responses to self-promotion. *Psychological science*, 26(6):903–914.

John R Searle, Ferenc Kiefer, Manfred Bierwisch, et al. 1980. *Speech act theory and pragmatics*, volume 10. Springer.

Ovul Sezer, Francesca Gino, and Michael I Norton. 2018. Humblebragging: A distinct—and ineffective—self-presentation strategy. *Journal of Personality and Social Psychology*, 114(1):52.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Tannen. 1990. Women and men in conversation. *New York: William Morrow*.

Roos Vonk. 2002. Self-serving interpretations of flattery: Why ingratiation works. *Journal of personality and social psychology*, 82(4):515.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. 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*.

Zhijun Xu, Siyu Yuan, Lingjie Chen, and Deqing Yang. 2024. "a good pun is its own reword": Can large language models understand puns? In *Proceedings of*

the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024, pages 11766–11782. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Bragging-related Tasks Prompt

A.1 Prompt for Bragging Recognition

To evaluate LLM performance on the Bragging Recognition task, we employed two carefully constructed prompt variations, depicted in Figures 3 and 4. These prompts were designed to test model sensitivity to the types of examples provided in the few-shot learning context.

Specifically, the prompt in Figure 3 is *bragging-oriented*, featuring only examples of bragging statements and their rationales. In contrast, the prompt in Figure 4 is *non-bragging-oriented*, exclusively containing examples of non-bragging statements. Apart from these differing example sets, both prompts provide the same definition of bragging, instruct the model to classify input text, and require a CoT-based reasoning step prior to the JSON output. This dual-prompt strategy allows us to measure the degree to which LLMs' classification decisions are influenced by the immediate exemplary context, providing insights into their robustness and generalization capabilities for this task.

/* Definition */
Bragging refers to the act of excessively boasting or speaking highly of one's own achievements, qualities, or possessions. It often involves exaggerating or expressing overconfidence. Please read the following sentences and determine if they exhibit bragging behavior.

/* Instruction */
Determine whether the given Text is a bragging. Give your reasons first, then make your final decision clearly. You must output the current status in a parsable JSON format. An example output looks like "Reason": "XXX", "Choice": "The given text is a XXX"

/* 10 Examples */
example: Text: I'm so exhausted from all these award ceremonies I have to attend. Output:"Reason": "The speaker is pretending to complain about being tired but is actually drawing attention to the fact that they are attending many award ceremonies, implying they are successful and in-demand. This is a common and relatively acceptable form of bragging in this context.", "Choice": "The Given test is a bragging."

Text: My phone is constantly buzzing with interview offers. Output:"Reason": "The speaker is subtly highlighting their professional success and desirability in the job market by mentioning the frequent interview offers they receive. While it appears to be a straightforward statement, it serves as an indirect form of bragging, showcasing their in-demand status and appealing career opportunities. This approach is commonly used and generally accepted in professional and social contexts.", "Choice": "The Given test is a bragging."

Text: My income is higher than all of yours combined, and I don't even need to work to live comfortably. Output:"Reason": "The speaker is overtly boasting about their wealth and financial independence, emphasizing that they earn more than others combined and no longer need to work for a comfortable lifestyle. This approach is unlikely to be well-received in most social or professional contexts.", "Choice": "The Given test is a bragging."
...

Figure 3: This prompt is biased towards bragging.

A.2 Prompt for Bragging Explanation

For the Bragging Explanation task, which assesses an LLM's ability to understand and articulate the social context and motivations behind bragging statements, we employed a carefully crafted prompting strategy. The prompts utilized a few-shot learning approach, providing models with a definition of bragging, explicit instructions, and, importantly, several examples of bragging statements paired with their detailed explanations.

Unlike the prompts for Bragging Recognition that included both bragging and non-bragging examples to test classification, the examples for this task were exclusively instances of bragging. This was intended to focus the LLM on the generation of coherent and insightful explanations. The prompt implicitly guided the models towards a Chain-of-Thought process by structuring the task such that the reasoning or justification for the bragging nature of the input text was the primary output. An illustrative example of the prompt structure is shown in Figure 5.

A.3 Prompt for Bragging Generation

For *Prompt-driven Bragging Generation* (Scenario 1 of the Bragging Generation task), which requires LLMs to synthesize bragging statements

/* Definition */
Bragging refers to the act of excessively boasting or speaking highly of one's own achievements, qualities, or possessions. It often involves exaggerating or expressing overconfidence. Please read the following sentences and determine if they exhibit bragging behavior.

/* Instruction */
Determine whether the given Text is a non-bragging. Give your reasons first, then make your final decision clearly. You must output the current status in a parsable JSON format. An example output looks like "Reason": "XXX", "Choice": "The given text is a XXX"

/* 10 Examples */
Text: I have a great workout today. Output:"Reason": "It is not considered bragging because it is a simple statement of fact about the speaker's personal experience.", "Choice": "The Given test is a non-bragging."

Text: I've been learning a lot from this book I'm reading. Output:"Reason": "The speaker is sharing a personal experience of learning without boasting or elevating themselves in relation to others. It's simply a neutral observation.", "Choice": "The Given test is a non-bragging."

Text: I've been working on improving my running time, and it's been going well. Output:"Reason": "The statement focuses on personal progress rather than on any comparison or desire for validation. It's a neutral remark about self-improvement.", "Choice": "The Given test is a non-bragging."

Figure 4: This prompt is biased towards non-bragging.

/* Definition */
Bragging refers to the act of excessively boasting or speaking highly of one's own achievements, qualities, or possessions. It often involves exaggerating or expressing overconfidence.

/* Instruction */
Determine why the given Text is a bragging. Give your reasons, you must output the current status in a parsable JSON format. An example output looks like "Reason": "XXX"

/* 10 Examples */
Text: I'm so exhausted from all these award ceremonies I have to attend. Output:"Reason": "The speaker is pretending to complain about being tired but is actually drawing attention to the fact that they are attending many award ceremonies, implying they are successful and in-demand. This is a common and relatively acceptable form of bragging in this context."

Text: My phone is constantly buzzing with interview offers. Output:"Reason": "The speaker is subtly highlighting their professional success and desirability in the job market by mentioning the frequent interview offers they receive. While it appears to be a straightforward statement, it serves as an indirect form of bragging, showcasing their in-demand status and appealing career opportunities. This approach is commonly used and generally accepted in professional and social contexts."

Text: My income is higher than all of yours combined, and I don't even need to work to live comfortably. Output:"Reason": "The speaker is overtly boasting about their wealth and financial independence, emphasizing that they earn more than others combined and no longer need to work for a comfortable lifestyle. This approach is unlikely to be well-received in most social or professional contexts."

Figure 5: The prompts for bragging explanation are similar to those for bragging recognition. However, for the explanation task, we only provide sentences that are classified as bragging.

You are tasked with generating a bragging sentence based on a given social context and speaker's intention. Your goal is to create a realistic, boastful statement that fits the provided scenario.

Here are the details for this task:
POTENTIAL_SOCIAL_CONTEXT
SPEAKERS_INTENTION
Instructions: 1. Carefully analyze the provided social context and speaker's intention. 2. Put yourself in the mindset of someone who wants to brag about their accomplishments or abilities. 3. Craft a sentence that clearly expresses a boast while considering the following guidelines: a. Ensure the sentence aligns with the described social context. b. Incorporate elements that reflect the speaker's intention. c. Include a subtle attempt to downplay the boast through self-awareness, if appropriate. d. Make the sentence sound natural and fitting for the given context. e. Keep the tone consistent with someone sharing their achievements or abilities.
Before providing your final output, wrap your thought process in <bragging_analysis> tags:
1. Identify key elements: - List important aspects of the social context - Note crucial points from the speaker's intention
2. Brainstorm bragging ideas: - Generate at least 3 potential bragging statements - For each idea, evaluate how well it fits the criteria (context, intention, subtlety, naturalness, tone)
3. Select and refine the best bragging sentence: - Choose the idea that best meets all criteria - Refine the sentence to improve its effectiveness
Remember to fully embody the role of someone who wants to brag, while still maintaining awareness of the social context and the speaker's specific intentions.
After your thought process, provide your final output in the following format:
[Insert the generated bragging sentence here]
[Provide a brief explanation of how the sentence meets the given criteria and reflects the social context and speaker's intention]

Figure 6: Example prompt for the *Prompt-driven Bragging Generation* task, outlining the instructions for LLMs to generate contextually appropriate bragging statements. It includes guidelines for analysis, brainstorming, selection, and a structured output format incorporating a reasoning step.

from given contextual cues, a specific instructional prompt was designed. As shown in Figure 6, this prompt directs the LLM to embody a bragging persona and generate a statement fitting a specified social context and speaker's intention. Key elements of the prompt include detailed guidelines for the generation process, a requirement for an explicit reasoning step (the '<bragging_analysis>' block), and a structured format for the final output.

This detailed generation prompt is specific to Scenario 1. The second scenario, ***Responding to User Bragging***, focuses on the LLM's reactive capabilities; thus, it involves prompting the LLM to respond to an existing brag rather than generating one from scratch based on such detailed instructions.

B The Details of Re-annotation Process

To facilitate a more nuanced evaluation of LLMs' ability to comprehend bragging, we significantly enhance the publicly available dataset originally curated by Jin et al. (2022). Recognizing that bragging is a complex social act, deeply intertwined with context, intention, and social interactions, our re-annotation focus specifically on sentences identified as bragging. We meticulously add detailed information about four crucial aspects: Potential Social Context, Speaker's Intention, Desired Feedback, and Appropriateness. ***1) Potential Social Context:*** This captures the setting or circumstances surrounding the bragging statement, recognizing that the same utterance can be interpreted differently depending on where and when it is expressed. ***2) Speaker's Intention:*** This delves into the underlying motivations for bragging, such as seeking validation, impressing others, or expressing genuine excitement. ***3) Desired Feedback:*** This identifies the type of response or reaction the speaker likely seeks, revealing the social expectations associated with the act of bragging. ***4) Appropriateness:*** This assesses the social acceptability of the bragging statement in the given context, considering factors like cultural norms and the relationship between speaker and listener. For each bragging statement, we also provide a comprehensive explanation justifying its classification, analyzing the interplay of these four elements.

This meticulous re-annotation process was undertaken by three postgraduate annotators with expertise in social linguistics. They were rigorously trained to identify and analyze the subtle

interactions of bragging behavior, equipped with detailed guidelines and illustrative examples to foster consistency in their judgments. Acknowledging the inherent subjectivity in human annotation, we implemented a two-tiered quality control process. First, annotators manually cross-checked a subset of their annotations to identify potential discrepancies. Second, we leveraged GPT-4 as an adjudication tool to harmonize annotations where significant disagreements arose, ensuring that each final annotation comprehensively incorporates the insights of all three annotators. The resulting dataset, encompassing the original bragging statements alongside our rich annotations and explanations, will be made publicly available upon publication of our research findings.

B.1 Annotation Training Protocol.

Three postgraduate annotators with a background in social linguistics and NLP were involved in the re-annotation. Prior to the main annotation task, they underwent a dedicated training phase which included:

1. **Guideline Comprehension:** A detailed annotation manual was developed, defining the four key dimensions for each bragging statement: Potential Social Context, Speaker's Intention, Desired Feedback, and Appropriateness. This manual included numerous examples to illustrate these dimensions and clarify potential ambiguities.

2. **Pilot Annotation and Calibration:** Annotators independently annotated a common pilot set of 50 bragging statements. Subsequent group discussions were held to resolve discrepancies, refine interpretations of the guidelines, and achieve a shared understanding. Inter-Annotator Agreement (IAA) was calculated on a subset after this phase to ensure reliability (Fleiss' Kappa > 0.85).

3. **Iterative Feedback:** Throughout the annotation process, regular meetings facilitated discussion of challenging cases and maintained consistency.

B.2 Multi-Stage Annotation Workflow with GPT-4 Assistance.

To clarify the precise role of LLMs and the primacy of human judgment, the annotation harmonization and finalization process followed these distinct stages:

1. **Initial Independent Human Annotation:** As described above, each of the three trained annotators independently provided annotations for all four dimensions for each bragging statement.

2. **GPT-4 Assisted Consistency Check:** The independently produced annotations for each statement were then presented to GPT-4 (gpt-4-0613). GPT-4 was prompted to compare the annotations across the three annotators for each dimension and highlight any inconsistencies or disagreements in labeling or rationale. It was specifically instructed not to provide a "correct" answer but to identify points of divergence.

3. **Human Review and Consensus-Driven Decision-Making:** All disagreements flagged by GPT-4, or any instances where annotators had initially expressed low confidence, were subjected to a thorough manual review by the three human annotators.

   • The annotators discussed the differing perspectives, referring back to the annotation guidelines and examples.

   • For some cases, GPT-4's output from the consistency check (which might articulate the nature of the disagreement) was used as a starting point for discussion, but never as the deciding factor.

   • A consensus was reached through this human deliberation. If a robust consensus could not be achieved for a particular statement even after discussion, that data point was excluded from the datasets used for the explanation and generation tasks to maintain high data quality.

4. **GPT-4 Assisted Summary Generation (for Consistent Annotations):** Once a consensus was reached for a bragging statement (either from initial agreement or after the human adjudication process), GPT-4 was then prompted to assist in drafting a concise summary explanation that encapsulated the agreed-upon Potential Social Context, Speaker's Intention, Desired Feedback, and Appropriateness, along with the overall rationale.

5. **Final Human Verification and Refinement:** Crucially, every summary description drafted

with GPT-4's assistance underwent a final meticulous review by at least two human annotators. Annotators verified the accuracy, completeness, and nuance of the summary, ensuring it faithfully represented the human-derived consensus. Edits were made as necessary. No annotation was considered final without this explicit human approval.

This detailed, multi-stage process, with human judgment at its core for all critical decision points, was designed to mitigate potential biases from relying solely on LLM interpretations and to ensure the final annotated dataset accurately reflects nuanced human understanding of bragging behavior.

B.3 Illustrative Example of the Process

To further illustrate this workflow, consider the following example:
Text: "Just closed the biggest deal in company history! Celebratory drinks tonight, my treat!"

1. ***Initial Human Annotations:***

   • *Annotator A:*

   Speaker's Intention: Primarily "expressing excitement and celebrating achievement."

   Appropriateness: "Relatively appropriate in a workplace setting, but slightly boastful."

   • *Annotator B:*

   Speaker's Intention: Includes not only "celebrating achievement" but also "seeking colleague recognition and praise."

   Appropriateness: "Appropriate in informal settings, but slightly inappropriate in formal work groups."

2. ***GPT-4 Assisted Consistency Check and Analysis:*** Annotators A and B show subtle differences in interpreting 'Appropriateness,' particularly concerning the distinction between 'formal' vs. 'informal' workplace settings. The phrase 'Celebratory drinks tonight, my treat!' implies an informal celebratory atmosphere, potentially making 'appropriate in informal settings' a more fitting interpretation. Annotator B also identifies 'seeking colleague recognition' as part of the intention, which is a plausible inference from the public announcement and invitation.

3. ***Human Review and Final Decision-Making (Post-Discussion):*** Based on GPT-4's analysis, Annotators A and B re-examined the text and their rationales, eventually reaching consensus:

   • **Speaker's Intention:** Celebrating achievement, seeking colleague recognition and praise (adopting GPT-4's suggestion)

   • **Appropriateness:** Appropriate in informal workplace settings, but slightly inappropriate in formal work groups (adopting Annotator B's judgment, inspired by GPT-4's analysis of 'informal celebratory atmosphere')

4. ***Final Annotation Summary (GPT-4 Assisted Generation, Human Verified):*** The speaker celebrates personal achievement by announcing closing the biggest deal in company history and inviting colleagues for celebratory drinks, implying seeking colleague recognition and congratulations. This bragging is relatively acceptable in informal workplace settings or teams, but may be considered less humble in more formal settings.

B.4 The Annotated Examples

Here are some of the results we annotated:
Example 1:

> **Text:** I have this one stock that has slowly been going up . No jumps or dips just slowly but surely.
> **Potential Social Context:** Casual conversation or social media discussion about personal investments, where financial success stories are shared.
> **Speakers Intention:** Demonstrate investment expertise while maintaining an appearance of modesty through emphasis on steady, rather than dramatic, gains.
> **Desired Feedback:** Recognition of their investment wisdom and congratulations on their success, while being perceived as a measured and thoughtful investor.
> **Appropriateness:** Relatively acceptable due to understated tone, though discussing personal financial success can still be sensitive in many social contexts.
> **Bragging Explanation:** The speaker indirectly boasts about their successful investment while using the 'slow but steady' framing to appear humble and wise rather than overtly boastful.

Example 2:

> **Text:** @USER I 'm 35 and nobody who sees my id can believe it .
> **Potential Social Context:** Social media platform where discussions about appearance and age are common, likely in response to or seeking validation from peers.
> **Speakers Intention:** To highlight their youthful

appearance while attempting to disguise the boast through a seemingly innocent observation about ID verification.
Desired Feedback: Recognition of their investment wisdom and congratulations on their success, while being perceived as a measured and thoughtful investor.
Appropriateness: Appropriate but obvious vanity, though somewhat softened by self-aware use of #humblebrag.
Bragging Explanation: The speaker deliberately draws attention to their supposedly youthful appearance while framing it as an observation from others, using the #humblebrag tag to acknowledge yet justify the self-promotion.

Example 3:

Text: Being stuck in the house only made me do more research on financial things and I 'm proud I did.
Potential Social Context: Social media post during or after COVID-19 lockdown period, where people commonly shared their lockdown activities and achievements.
Speakers Intention: To differentiate themselves from others by highlighting their productive use of lockdown time for self-improvement in financial literacy.
Desired Feedback: Recognition and admiration for their perceived wisdom and initiative in using confined time for self-improvement.
Appropriateness: Moderately acceptable as it's framed as personal growth, though it carries an implicit criticism of those who didn't use lockdown time.
Bragging Explanation: The statement subtly implies superiority by presenting forced confinement as an opportunity seized, while using passive voice and emojis to soften the self-promotion.

C Metrics for Bragging Explanation

Our selection of evaluation metrics is fundamentally driven by the comprehensive annotation process we undertook. Since our re-annotation of the dataset meticulously identified the Potential Social Context, Speaker's Intention, Desired Feedback, and Appropriateness for each bragging statement, these elements naturally became the core criteria for assessing the quality of LLM-generated explanations. We suppose that a high-quality explanation, particularly when created through a Chain-of-Thought (CoT) approach, should mirror the human reasoning process used during annotation. Therefore, the LLMs' explanations should demonstrably incorporate these four key aspects to show a complete understanding of the bragging utterance.

Our evaluation methodology focuses on quantifying the presence of these elements within the LLMs' explanations. For each explanation, we determine whether each of the four elements is explicitly mentioned and correctly identified. The final assessment is based on the percentage of explanations that successfully incorporate each of these crucial aspects. We contend that this approach directly measures the LLMs' ability to grasp the same nuanced understanding of bragging that informed our annotation process.

While the rationale for choosing these specific metrics is further elaborated in Appendix B, it is important to emphasize here that their selection is intrinsically linked to the rich information captured in our annotated dataset. By focusing on these elements, we aim to evaluate not just the surface-level accuracy of the explanations but also the depth of the LLMs' comprehension of the underlying social dynamics of bragging.

D Metrics for Bragging Generation

The overarching goal of this section is to assess how well LLMs can generate bragging-related content and respond to it in a socially appropriate manner. This requires evaluating both the quality of the generated bragging itself and the quality of the interaction when the LLM acts as a listener.

D.1 Scenario 1: Prompt-driven Bragging Generation

This scenario aims to evaluate the LLM's ability to produce bragging statements that are relevant to a given prompt, appropriate for a specified social context, and effective in achieving a social goal. The metrics are chosen to capture these different facets of bragging generation.

We have manually evaluated 200 generated sentences, each assessed by three independent human annotators, using the following criteria. Each metric is scored on a scale from 0 to 10, where 0 represents the lowest and 10 represents the highest. Each metric is described in terms of its purpose and the evaluation process.

Bragging Success: This metric directly assesses whether the generated text is perceived as boasting or self-promotion. It serves as a fundamental check; if a statement is not perceived as bragging, the subsequent metrics lose their relevance. Human annotators rate the degree to which the statement conveys a sense of self-aggrandizement or self-promotion. Scores range from 0 (not perceived as bragging at all) to 10 (clearly perceived as bragging).

Complied Social Context Human annotators judge the appropriateness of the generated brag-

ging statement with respect to the provided social context. Scores range from 0 (completely inappropriate) to 10 (completely appropriate).

Social Goal Achievement Bragging is highly context-dependent. This metric ensures that the generated statement aligns with the specified social context that was provided during the generation phase. Human annotators judge how well the generated bragging statement aligns with the provided social context. Scores range from 0 (completely inconsistent with the provided social context) to 10 (completely consistent with the provided social context).

Bragging Intensity Tragging often serves a particular social goal. This metric assesses whether the LLM generates a statement that effectively supports the social goal that was specified during the generation phase. Human annotators rate the effectiveness of the generated bragging statement in achieving the implied social goal, as defined during the generation phase. Scores range from 0 (not effective at all in achieving the defined goal) to 10 (very effective in achieving the defined goal).

To complement human evaluation, we also employed an automatic metric to assess the use of humor in the generated bragging statements, recognizing that humor often plays a key role in making bragging more palatable, especially in informal contexts. We utilized a computational model, as described in Baranov et al. (2023), to measure the presence and extent of humor. This model provides a confidence score indicating the likelihood of a given text being humorous. We adopted this confidence score as a proxy for the degree of humor, hereafter referred to as **huMor**, providing an objective and scalable way to evaluate this important stylistic element of successful bragging.

D.2 Scenario 2: Responding to User Bragging

This scenario evaluates the LLM's ability to act as a socially intelligent listener when presented with a bragging statement (generated in Scenario 1 and provided as input for this stage). The metrics are designed to assess the appropriateness and emotional alignment of the LLM's response. We employed human evaluation for two aspects, and an automatic metric for sentiment analysis.

Preachiness Intensity (PI) Responding to bragging with a preachy or moralizing tone is generally considered socially inappropriate. This metric ensures the LLM avoids sounding judgmental or didactic. Human annotators rate the degree to which the LLM's response is overly didactic or moralizing. Scores are on a scale from 0 to 5, where 0 indicates no preachiness and 5 indicates a highly preachy response.

Sycophancy Intensity (SI) While admitting to bragging is important, excessive flattery or sycophantic flattery may be false or offensive. This metric ensures the LLM strikes a balance between politeness and social engagement. Human annotators rate the extent to which the LLM's response is excessively flattering. Scores are on a scale from 0 to 5, where 0 indicates no sycophancy and 5 indicates a highly sycophantic response.

Complementing human evaluations, we incorporated an automatic metric to gauge the sentiment alignment between the generated bragging statement and the LLM's response. Recognizing that an appropriate response to bragging often acknowledges or validates the positive emotion conveyed by the bragger, we employed a sentiment analysis model (Sanh, 2019), fine-tuned on the SST-2 dataset (Socher et al., 2013), to calculate a **Sen**timent **Gap**. This metric quantifies the difference in sentiment polarity scores, as determined by the model's confidence in classifying each statement as positive or negative, between the original bragging statement (generated in Scenario 1) and the LLM's subsequent response. A smaller sentiment gap indicates better alignment, suggesting the response appropriately reflects the positive sentiment of the bragging.