"My life is miserable, have to sign 500 autographs everyday": Exposing Humblebragging, the Brags in Disguise

Sharath Naganna*, Saprativa Bhattacharjee*, Biplab Banerjee, Pushpak Bhattacharyya
Indian Institute of Technology Bombay, Mumbai, India
{sharathhn, saprativa, pb}@cse.iitb.ac.in, bbanerjee@iitb.ac.in

Abstract

Humblebragging is a phenomenon in which individuals present self-promotional statements under the guise of modesty or complaints. For example, a statement like, "*Ugh, I can't believe I got promoted to lead the entire team. So stressful!*", subtly highlights an achievement while pretending to be complaining. Detecting humblebragging is important for machines to better understand the nuances of human language, especially in tasks like sentiment analysis and intent recognition. However, this topic has not yet been studied in computational linguistics. For the first time, we introduce the task of automatically detecting humblebragging in text. We formalize the task by proposing a 4-tuple definition of humblebragging and evaluate machine learning, deep learning, and large language models (LLMs) on this task, comparing their performance with humans. We also create and release a dataset called HB-24, containing 3,340 humblebrags generated using GPT-4o. Our experiments show that detecting humblebragging is non-trivial, even for humans. Our best model achieves an F1-score of 0.88. This work lays the foundation for further exploration of this nuanced linguistic phenomenon and its integration into broader natural language understanding systems.

1 Introduction

Humblebragging is a nuanced socio-linguistic phenomenon in which individuals subtly boast about their achievements, possessions, or qualities while disguising their self-promotion with expressions of complaint or modesty. The term was coined by American comedian Harris Wittels in 2010, who later published a book on the phenomenon titled *Humblebrag: The Art of False Modesty* (Wittels, 2012). A few examples of humblebrags are provided in Table 1.

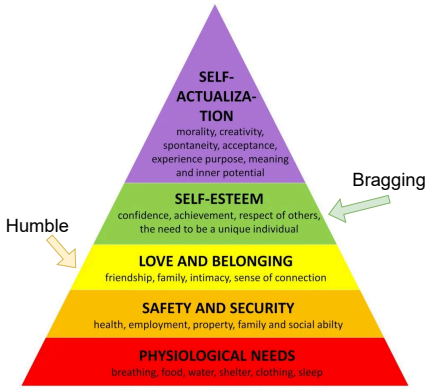


Figure 1: Why do people resort to humblebragging? The answer lies in the Maslow's Hierarchy of Needs[1]. Humblebragging satisfies needs of belonging and self-esteem simultaneously.

To address why people resort to humblebragging, Trivedi and Srinivas (2019) analyzed its rising popularity, particularly on social media, from the dual perspectives of need theories (*Maslow's Hierarchy of Needs*, Maslow and Lewis 1987; see Figure 1) and the *Hubris Hypothesis* (Hoorens et al., 2012). Humblebragging allows a person to satisfy the dual needs of belonging (level 3) through humility; and self-esteem (level 4) through bragging. Moreover, the hubris hypothesis states that addressees prefer implicit self-superiority claims over explicit ones. Also, straightforward bragging violates the maxims of modesty (Leech, 1983) and self-denigration (Gu, 1990), as noted by Zuo (2023). Humblebragging thus serves as a subtle way to convey positive information about oneself, often through expressions of modesty or complaint. Even though it is commonly observed on social media (Twitter, Reddit, and Instagram), many people also use it in real-life conversations (Wittels, 2012; Sezer et al., 2018; Trivedi and Srinivas, 2019). Although humblebragging has been recognized socially and culturally, it

*Equal contribution.

[1]Image source: https://www.simplypsychology.org/maslow.html

Findings of the Association for Computational Linguistics: ACL 2025, pages 3838–3858

remains unexplored through the lens of computational linguistics. Evidence from search trends, social media behavior, and historical usage suggests that humblebragging is a persistent and recognizable phenomenon, even when not explicitly labeled. We provide supporting statistics and examples in Appendix A.

Humblebragging is closely related to other forms of figurative language, such as *sarcasm* (Gibbs, 1986) and *irony* (Garmendia, 2018), which also rely on *verbal incongruity*, a contrast between what is said and what is meant. While irony typically contrasts expectation with reality and sarcasm adds a mocking tone, humblebragging uniquely conceals self-promotion within a modest or complaining remark. A detailed comparison of these differences is provided in Appendix B. Although extensive research exists on computational modeling of sarcasm (Bhattacharyya and Joshi, 2017; Riloff et al., 2013; Cai et al., 2019) and irony (Zeng and Li, 2022; Van Hee et al., 2018; Barbieri and Saggion, 2014), to the best of our knowledge, this work is the first to explore humblebragging in computational linguistics.

Detecting humblebragging automatically is crucial for a nuanced understanding and processing of human language. This is necessary for enhancing accuracy of tasks such as sentiment analysis (Zhang et al., 2024), intent recognition (Lamanov et al., 2022), emotion recognition (Li et al., 2022) and dialogue understanding (Gao et al., 2024). In applications like social media monitoring and customer feedback analysis, it helps differentiate genuine complaints (Singh et al., 2023) from brags disguised as complaints. Moreover, this capability is also valuable for researchers in the humanities and social sciences.

In this paper, we formally introduce the task of humblebragging detection and present a curated dataset that combines existing resources with synthetic data generated using GPT-4o. By addressing this previously unexplored area, our work bridges the gap between computational linguistics and other disciplines in the study of humblebragging.

Our contributions are:

1. Introduction of the task of automatic humblebragging detection from text to the computational linguistics community by proposing a 4-tuple definition of humblebragging for streamlining its processing (section 3 and section 5).

2. Benchmarking of various machine learning, deep learning, and state-of-the-art large language model techniques on the task of automatic humblebragging detection from text (section 6 and section 7).

3. Release of a new dataset named HB-24 on humblebragging detection, containing 3,340 humblebrags, to enable further research on the task (section 4).

2 Related Work

In Psychology and Other Disciplines Humblebragging has been extensively studied in psychology. Sezer et al. (2018) examined its effects on audiences, showing that humblebragging is ubiquitous in daily interactions, with 70% of humblebrags falling into the complaint-masked variety. Trivedi and Srinivas (2019) explained the widespread use of humblebragging through the dual perspectives of need theories and the hubris hypothesis, and also provided a contextual framework for understanding it. Other notable studies include Sezer et al. (2015); Vranka et al. (2017); Luo and Hancock (2020). Beyond psychology, humblebragging has been explored in disciplines like tourism research (Chen et al., 2020; Yan et al., 2024), pragmatics (Lin and Chen, 2022; Zuo, 2023; Han et al., 2024), and advertising (Paramita and Septianto, 2021).

Sarcasm and Irony Both *sarcasm* and *irony* have been extensively studied in computational linguistics over the past two decades. Joshi et al. (2015) demonstrated how incongruences can enhance sarcasm detection, while Joshi et al. (2017a) categorized detection methods, benchmark datasets, and evaluation metrics. Recently, Gole et al. (2023) explored the use of large language models for sarcasm detection. Beyond detection, Joshi et al. (2017b) proposed a hybrid rule-based and statistical approach for identifying sarcasm targets, which was later complemented by transformer-based methods such as BERT, as demonstrated by Parameswaran et al. (2021). For irony, Zeng and Li (2022) provided a comprehensive survey on computational approaches. Hernández-Farías et al. (2015) evaluated traditional machine learning models for irony detection using sentiment scores, while Wen et al. (2023) introduced the *Retrieval–Detection Method for Verbal*

Humblebrags	Mask type
I can't believe they'd give an idiot like me a phd lol	Modesty
Being in demand means disappointing 95% of people 95% of the time. I have yet to learn how to overcome this.	Modesty
For the 3rd time in 3 years I've been asked to speak at Harvard, but I've yet to speak at my alma mater. What's a girl gotta do @MarquetteU?	Complaint
Will Twitter be available for me in Paris, milan, or the Maldives? I hope so bc it won't in hong Kong or Singapore	Complaint

Table 1: Examples of humblebrags. Each instance of a humblebrag consists of a brag masked by either complaint or modesty. The brags are in red while the masks are in blue.

Irony (RDVI), leveraging open-domain resources for enriched detection.

Bragging and Humility A closely related area of research is the detection and processing of *bragging* (Alfano and Robinson, 2014) and *humility* (Snow, 1995) as standalone tasks. Jin et al. (2022) introduced bragging classification to the computational linguistics community and released a public dataset, while Jin et al. (2024) conducted a large-scale study of bragging behavior on Twitter. For humility, Guo et al. (2024) explored LLM-based techniques for measuring humility in social media posts. Additionally, Danescu-Niculescu-Mizil et al. (2013); Firdaus et al. (2022); Srinivasan and Choi (2022) have examined *politeness* in computational contexts.

Synthetic Data Generation The language generation capabilities of LLMs have created opportunities for generating synthetic data. Long et al. (2024) provide a comprehensive survey of synthetic data generation, curation, and evaluation, while Li et al. (2023) explore the potential and limitations of using LLMs for this purpose. Synthetic data generated by language models has been applied to various text classification tasks (Chung et al., 2023; Sahu et al., 2022; Ye et al., 2022; Yoo et al., 2021).

Classification as Generation Finally, there has been a growing trend towards performing classification tasks by posing them as generation tasks. This approach to text classification is particularly relevant for leveraging decoder-based large language models in classification settings, where text generation mechanisms complement traditional methods. For instance, LLMs have been employed as zero-shot (Gretz et al., 2023) and few-shot (Mirza et al., 2024) text classifiers. Moreover, Saunshi et al. (2021) provides mathematical insights into modeling classification tasks as text completion tasks.

3 Formulation of the Humblebragging Definition

In this section, we define and derive our proposed framework for humblebragging.

3.1 Formal Definition of Humblebragging

We define humblebragging as a 4-tuple to systematically capture its key components and underlying structure:

$$HB = \langle B, BT, HM, MT \rangle \quad (1)$$

where:

- B: **Brag** – The segment of the text that explicitly conveys the act of bragging.
- BT: **Brag Theme** – The overarching theme or specific category of the brag embedded within the statement. Categories are listed in Appendix C.
- HM: **Humble Mask** – The segment of the text that adopts a modest or complaining tone to obscure or mitigate the act of bragging.
- MT: **Mask Type** – Specifies whether the humble mask adopts a modest tone or a complaining approach.

For instance, in the following statement:

"Ugh, I can't believe I got promoted to lead the entire team. So stressful!"

- **B**: *"I can't believe I got promoted to lead the entire team."*;
- **BT**: Performance at work;
- **HM**: *"Ugh,"* and *"So stressful!"*;
- **MT**: Complaint.

3.2 Derivation of 4-tuple Definition

Our 4-tuple definition is adapted from the 6-tuple framework of sarcasm (Ivanko and Pexman, 2003): ⟨**Context (C), Utterance (u), Literal Proposition (p), Intended Proposition (p'), Speaker (S), and Hearer (H)**⟩. In sarcasm, p conveys a surface-level meaning that contrasts with p', creating incongruity. In humblebragging, the same phenomenon

		Humblebrag	Non-humblebrag
Train	Samples	3340	5431
	Min #words	6	1
	Max #words	47	68
	Avg #words	15.98	16.41
Test	Samples	558	576
	Min #words	1	6
	Max #words	70	47
	Avg #words	19.55	17.5

Table 2: Dataset statistics. Min, Max and Avg refer to minimum, maximum and average respectively.

is achieved through a **Brag (B)** and a **Humble Mask (HM)**, where HM corresponds to p, presenting a modest or complaining front, while B aligns with p', subtly revealing the self-promotion. The Context (C) maps to the **Brag Theme (BT)**, which categorizes the nature of the brag (e.g., achievements, wealth, intelligence). Additionally, the **Mask Type (MT)** incorporates the classifications into modesty or complaint, further refining the nature of the humblebrag.

Unlike sarcasm, where the Speaker (S) directs an utterance toward a Hearer (H) who must infer the intended meaning, humblebragging often lacks a specific hearer. It is frequently self-directed or broadcasted to a broad audience, making S and H unnecessary in this framework. This adaptation preserves the dual-layered meaning from sarcasm while formalizing humblebragging as a strategic blend of self-effacement and self-promotion.

4 Dataset

As there were no existing datasets for the task of *humblebragging detection*, we propose *HB-24*[2], a well-balanced collection of humblebrag and non-humblebrag texts, comprising both human-written and synthetic samples. Due to the limited availability of quality humblebrags, we leverage the capabilities of large language models to generate human-like examples, augmenting the existing data and enhancing the corpus for training classification models. The synthetic data is used for training, while the trained model's performance is evaluated on human-written samples. In other words, the training set is composed of synthetic humblebrags while the test set consists of human-written humblebrags. The non-humblebrags are all human-written. Table 2 presents the dataset statistics.

[2]Available at https://github.com/SharathHN/HB-24

Prompt Type	#Samples
General Prompt	1100
Prompt with Themes	1304
Few-Shot with Themes	936
Total	3340

Table 3: Prompt type and the number of samples.

4.1 Human-Written Humblebrags

Wittels (2012) presents a curated collection of high-quality humblebrag texts, categorized into themes such as wealth, first-class travel, workplace achievements, celebrity status, and more. These tweets form the positive class within our test set.

4.2 Synthetic Humblebrags

The humblebrags in the training set consist entirely of synthetic tweets generated using GPT-4o through zero-shot and few-shot prompting. The prompt template follows a format similar to that of Li et al. (2023). In the following sections, we discuss the prompts used for generating synthetic humblebrags.

4.2.1 Zero-Shot Prompts

In the zero-shot generation setup, we used two types of prompts. In the *General Prompt*, we did not explicitly define humblebragging; instead, we asked the model to generate tweets that subtly mention various achievements. In the *Prompt with Themes*, we provided a formal definition of humblebragging along with the themes (Appendix C) outlined in Wittels (2012). The prompts are provided in Appendix D.

4.2.2 Few-Shot Prompts

In the few-shot prompt setup, we modified the *Prompt with Themes* to include a few examples from each theme. We experimented with varying numbers of examples, starting with one and going up to five, and observed that increasing the number of examples did not improve the generation quality. Consequently, we settled on three examples per prompt for generating samples with few-shot prompts.

4.2.3 Post-Processing and Data Curation

After executing all three prompts, we generated a total of 11,000 synthetic samples containing humblebrags. Each sample was manually reviewed to assess its quality and relevance (see Appendix E for more details). From this pool, we filtered tweets

	Train	Test
Sarcasm	16%	12%
Humblebrag	38%	49%
Irony	15%	11%
Complaints	14%	10%
Neutral	14%	15%
Bragging	3%	4%

Table 4: Dataset composition.

from each prompt type, ensuring a balanced selection (see Table 3).

4.3 Non-Humblebrags

Humblebrags are often confused with sarcasm and irony, as all three involve an incongruence between the utterance and its intended meaning. To help the model distinguish these phenomena, we included sarcasm and irony as negative samples, alongside direct brags and straightforward complaints. Direct brags convey explicit self-promotion, while complaints reflect surface emotions often present in humblebrags. A more detailed discussion on the differences can be found in Appendix B.

Sarcastic samples were sourced from the SARC dataset (Khodak et al., 2018), and ironic ones from SemEval-2018 (Van Hee et al., 2018). Brags and complaints were taken from Jin et al. (2022) and Preoţiuc-Pietro et al. (2019), respectively. Neutral sentences, essential for improving class distinction (Koppel and Schler, 2006), were drawn from SemEval-2017's sentiment analysis task (Rosenthal et al., 2017). The dataset composition is shown in Table 4.

5 Methodology

We define two task setups: a standard binary classification task, and a sentence completion formulation where humblebrag detection is cast as a Yes/No question answering problem. This enables effective use of decoder-only language models for classification through natural language prompts.

5.1 Binary Classification for Humblebragging Detection

In the context of humblebragging detection, the task is to classify a given text x as either C_{HB} (Humblebragging) or $C_{\text{Non-HB}}$ (Non-humblebragging). The process involves generating text encodings from the input, which are then used for classification. For further details, we refer the reader to Appendix F.

5.2 Classification as a Sentence Completion Task with Yes/No Questions

Though decoder models are primarily designed for language generation tasks, their ability to predict the next token in a sequence makes them adaptable to various natural language understanding tasks, including classification. Humblebragging classification can be reformulated as a Yes/No question-answering task, where the model determines whether the input text contains humblebragging or not. This approach leverages the natural language understanding capabilities of pre-trained language models to classify text.

Framework The input text x is transformed into a prompt structured as

`<definition><question><x><answer>`

The model is given the `<question>` along with `<definition>` and `<x>` as the input prompt and is expected to generate a text completion for the `<answer>`. The LLM output in `<answer>` is then analyzed to determine whether it contains the required word.

- If `<answer>` contains **"Yes"**, the input is classified as y=1 (Humblebragging).
- If `<answer>` contains **"No"**, the input is classified as y = 0 (Non-humblebragging).

We evaluate this framework under two settings: **Z** (zero-shot) and **Z+D** (zero-shot + definition). In the **Z** setting, the definition is considered *null*, i.e., no external guidance is provided. In the **Z+D** setting, the 4-tuple definition is prepended to the input prompt to provide the model with additional context or instruction.

Example:

Input to LLM (Z+D setting):

`<definition>: HB = <B,BT,HM,MT>`

`<question>: Is the given text humblebragging or not? Answer in Yes or No only.`

`<x>: "Can someone tell the awards committee to chill? Running out of shelf space here!"`

Output from LLM:

`<answer>: Yes`

Classification: y=1

Detailed prompts for both settings are provided in Appendix G.

6 Experimental Setup

We conducted experiments with machine learning classifiers, encoder models, decoder models and compared the performance with those of human annotators.

6.1 Machine Learning Classifiers

We evaluated logistic regression and support vector machine (SVM) as simple machine learning based baselines to gauge the task difficulty.

6.2 Encoder Models

Two transformer-based (Vaswani et al., 2017) encoder models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), were evaluated on the task of humblebragging classification. The experiments utilized the `Adam` optimizer, and 5-fold cross-validation was employed for hyperparameter tuning.

6.3 Decoder Models

For humblebrag detection, decoder models were evaluated in zero-shot (Z) and zero-shot with definition (Z+D) settings, and were also fine-tuned using LoRA (F). In the Z setting, models classify statements as humblebrag or not using only the input. In Z+D, they leverage the 4-tuple definition to guide classification.

Further details about the settings, hyperparameters and human annotators can be found in Appendix H.

7 Results and Discussion

We present the results of all our experiments in Table 5. The table begins with the majority class baseline, which, in our case, involves predicting every sample as *non-humblebrag*.

7.1 Quantitative Analysis

Overall, the best-performing model in terms of F1-score is GPT-4o (0.88 F1), surpassing even the best human annotator (0.85 F1). We speculate that this may be due to the extensive linguistic and world knowledge these large-parameter models possess. Among the three human annotators, one performed significantly worse than the others, indicating that the task is non-trivial and can be challenging for some individuals. A detailed discussion on annotation inconsistency is provided in Appendix I. Moreover, two notable observations emerge from the results.

First, across all decoder models, the Z+D versions consistently outperform their Z counterparts, indicating that our 4-tuple definition effectively aids in detecting humblebragging. To further assess the specific contribution of our definition, we conducted two controlled experiments: (a) replacing the definition in the system prompt with random gibberish (Appendix J), and (b) substituting our 4-tuple definition with the 'textbook definition' (TD) by Wittels (2012) (Appendix K).

The results, summarized in Figure 2, clearly demonstrate that models prompted with our 4-tuple definition perform better than both the gibberish and textbook alternatives, reaffirming its utility in enhancing humblebrag classification.

Second, fine-tuning with our HB-24 dataset improved the F1-scores of the majority of the models. Both encoder models and three decoder models (Llama, Gemma, and Qwen) showed significant gains from fine-tuning. Interestingly, fine-tuned RoBERTa outperformed all 7–8 billion parameter decoder models except for Llama (F). This highlights the superior classification capabilities of encoder-only models when a high-quality dataset for fine-tuning is available.

We illustrate in Figure 3 Llama's progression from an F1-score of 0.66 in the zero-shot setting (Z) to an F1-score of 0.79 after fine-tuning (F), through confusion matrices. Llama (Z) primarily predicted the *yes* label for almost all samples. With our 4-tuple definition in Llama (Z+D), the model began to identify non-humblebrags, bringing more balance to the confusion matrix. After fine-tuning, Llama (F) became more proficient in identifying non-humblebrags while sacrificing some true positives. Confusion matrices of other models can be found in Appendix L.

Lastly, we observed anomalous behavior with Mistral and Vicuna, where fine-tuning led to lower F1-scores. Notably, Mistral's zero-shot performance already exceeded that of other models in its category, including the larger GPT-3.5 and any further fine tuning is resulting in catastrophic forgetting. In Vicuna's case, the fine-tuned model produced random texts and emojis and was extremely sensitive to slight prompt changes in the Z+D setting, requiring removal of the final sentence from the system prompt (Appendix G). Additional insights on performance degradation are provided in Appendix M.

Model	Accuracy	Precision	Recall	F1-Score
Baseline	0.51	0.25	0.50	0.34
Human 1	0.86	0.89	0.81	0.85
Human 2	0.84	0.86	0.81	0.84
Human 3	0.70	0.82	0.51	0.63
Average	0.80	0.86	0.71	0.77
Logistic Regression	0.59	0.68	0.58	0.53
SVM	0.62	0.72	0.61	0.56
BERT-Large-Uncased (F)	0.68	0.76	0.50	0.61
RoBERTa-Large (F)	0.78	**0.91**	0.62	0.74
GPT-4o (Z)	0.84	0.78	0.94	0.85
GPT-4o (Z+D)	**0.89**	**0.91**	0.85	**0.88**
GPT-3.5 (Z)	0.61	0.65	0.60	0.57
GPT-3.5 (Z+D)	0.75	0.76	0.75	0.75
Qwen2.5-7B-Instruct (Z)	0.64	0.82	0.35	0.49
Qwen2.5-7B-Instruct (Z+D)	0.71	0.85	0.50	0.63
Qwen2.5-7B-Instruct (F)	0.67	0.85	0.40	0.54
Mistral-7B-Instruct-v0.3 (Z)	0.60	0.55	0.96	0.70
Mistral-7B-Instruct-v0.3 (Z+D)	0.60	0.55	0.96	0.70
Llama-3.1-8B-Instruct (Z)	0.49	0.49	**0.99**	0.66
Llama-3.1-8B-Instruct (Z+D)	0.68	0.62	0.88	0.72
Llama-3.1-8B-Instruct (F)	0.81	0.87	0.72	0.79
Gemma-1.1-7b-it (Z)	0.57	0.57	0.57	0.57
Gemma-1.1-7b-it (Z+D)	0.56	0.53	0.83	0.65
Gemma-1.1-7b-it (F)	0.71	0.71	0.44	0.60
Vicuna-7b-v1.5 (Z)	0.55	0.60	0.28	0.38
Vicuna-7b-v1.5 (Z+D)	0.61	0.62	0.51	0.56

Table 5: Results of humblebragging classification. Z: zero-shot, Z+D: zero-shot with 4-tuple definition, F: fine-tuned. The best values are in bold.

7.2 Qualitative Analysis

In this section, we analyze cases of agreement and disagreement between human annotators and models.

All humans and models correctly classified the following as humblebrags, as the brags were easy to identify with a clear distinction between the brag and the mask segments:

> T1: *In the limo riding to airport. Sucks being alone though*
>
> T2: *just tried to pre-order my book. couldnt figure it out. did anyone try?*

For the following humblebrags, humans classified them as *no*, while models classified them as *yes*:

> T3: *I forget. What airport do u fly into to get to Maui?*
>
> T4: *I just had my first screaming girl encounter. She probably had me confused withsomeone else.*

In the case of T3, the disagreement could stem from a lack of knowledge regarding Maui as an exotic travel destination. On the other hand, the phrase "screaming girl" in T4 might not have been understood by the annotators due to cultural differences. Some cultures might interpret the phrase at its surface level without delving into its deeper meaning.

For the following humblebrag, humans said *yes*, while models said *no*:

> T5: *The CNN-LA green room is a cold and lonely place at 7 on a Sunday morning.Funnily enough, CNN LA green room a cold and lonely place at 10 on a Monday too.*

We hypothesize that the model might have been confused by the incongruity between "cold and lonely" and "Funnily enough," interpreting it as sarcasm instead of a humblebrag.

For the following non-humblebrag, both humans and models classified it as *yes*:

> T6: *i decided to become my own boss to have more free time.. now i have no time left whatsoever.*

T6 is a rare case where our assumption that humblebrags are not present in the datasets used to create our negative samples was violated. After encountering this example, we reviewed our dataset again to ensure no other such case exists.

For the following non-humblebrag, models said *no*, but humans said *yes*:

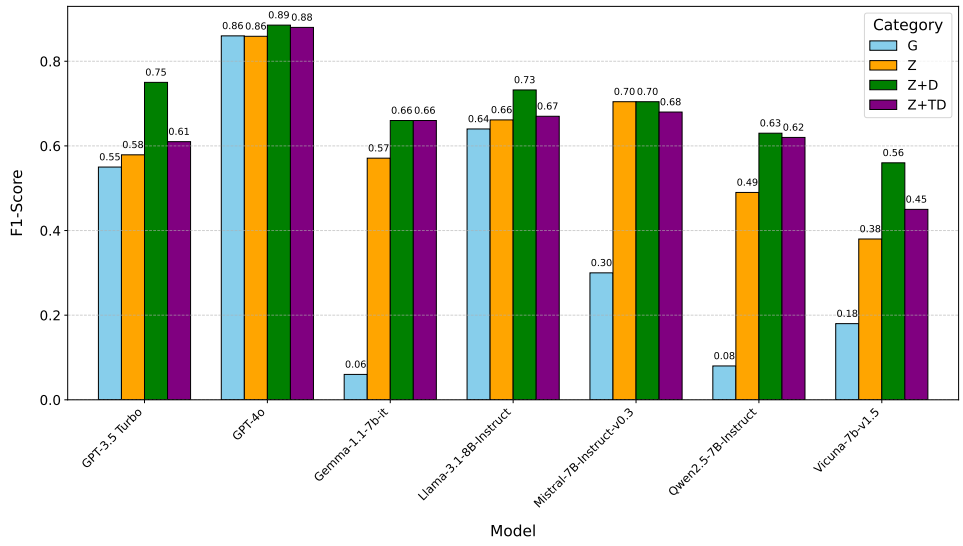
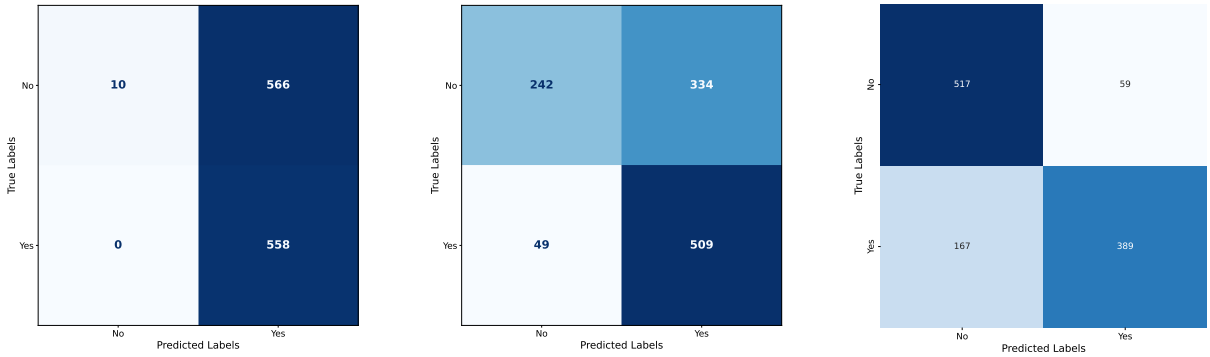


Figure 2: Comparison of F1-scores across scenarios for decoder models: G (Gibberish), Z (Zero-Shot), Z+D (Zero-Shot with our proposed 4-tuple definition of humblebragging), and Z+TD (Zero-Shot with the textbook definition of humblebragging). The Z+D setting achieves the highest F1-scores across all models, demonstrating the effectiveness of the proposed 4-tuple definition in capturing humblebragging nuances.



(a) Llama (Z) (b) Llama (Z+D) (c) Llama (F)

Figure 3: Confusion matrices for Llama Z vs Z+D vs F. Gradual improvement in the model can be observed from left to right.

T7: *After years of fumbling around , I have finally found a skin care product that works for me . Well , at least for now ?*

T8: *TheoCorleone david_maclellan Shit! I better shut my stupid girly mouth because im so concerned about what men might think of me.*

T7 and T8 represent classic cases of annotator bias, where annotators attempt to imagine a non-existent context and incorrectly classify the samples as belonging to the positive class. This bias arises because annotators, subconsciously influenced by their task, oversearch for humblebrags in the data as they are tasked to annotate for humblebragging detection.

While our qualitative analysis highlights the nuanced and often subjective nature of humblebrag detection, it also reveals patterns in how models and humans interpret brag-masking cues utilizing our 4-tuple framework. In the next subsection, we compare our framework with another plausible alternative to further reaffirm its suitability for the task.

7.3 4-tuple vs Sentiment Opposition Model

We pit our 4-tuple framework against the Sentiment Opposition Model (SOM) (Appendix N), which is based on detecting incongruity between surface sentiment and intended sentiment. For instance when a statement appears negative or modest but

Model	A	P	R	F1
Z+D	**0.68**	0.62	**0.88**	**0.72**
Z+SOM	0.66	**0.64**	0.74	0.68

Table 6: Comparison of Llama-3.1-8B-Instruct using the 4-tuple Definition (Z+D) vs. the Sentiment Opposition Model (SOM). A: Accuracy, P: Precision, R: Recall, F1: F1-score.

Model	A	P	R	F1
R-SST2	0.53	0.69	0.53	0.51
R-HBSC	**0.82**	**0.86**	**0.82**	**0.83**

Table 7: Comparison of vanilla RoBERTa-SST2 (R-SST2) with RoBerta-HBSC (R-HBSC). A: Accuracy, P: Precision, R: Recall, F1: F1-score.

actually conveys an underlying brag.

Table 6 indicates that the 4-tuple Definition (Z+D) outperforms the Sentiment Opposition Model (Z+SOM) in detecting humblebragging, achieving a higher recall and F1-score while maintaining comparable precision. This suggests that the 4-tuple framework more effectively captures both the brag and its masking component, leading to better overall detection. While SOM improves precision by reducing false positives, it sacrifices recall, making it less sensitive to subtle humblebraggings. The higher F1-score of the 4-tuple model confirms its better overall balance between precision and recall, making it a more robust approach for humblebragging detection.

7.4 Humblebragging Component Identification

We go a step further in testing the ability of models in identifying the brag and mask components in a given humblebrag by applying our definition. This structured interpretation of humblebragging text demonstrates the practical usefulness of our definition in capturing the nuances of humblebragging enabling models to distinguish the underlying brag from its masked presentation, making detection more accurate and interpretable. Appendix O showcases how different language models process humblebragging using this framework, highlighting variations in their ability to correctly segment and classify such statements. This analysis reinforces the effectiveness of our definition in improving both automated detection and human understanding of humblebragging in natural language.

7.5 Impact of Humblebragging Detection on Downstream Applications

To evaluate the utility of humblebragging detection in a downstream task, we conducted an intended-polarity classification experiment. From the HB-24 test set, we filtered out irony and neutral cases and defined gold labels as follows: humblebragging and bragging were labeled positive, while sarcasm and complaints were labeled negative.

As a baseline, we used RoBERTa-large fine-tuned on the SST-2 dataset (Socher et al., 2013), referred to as R-SST2. We then introduced a GPT-based classifier to detect humblebragging and sarcasm, adjusting the sentiment scores accordingly: +1 for humblebrags (to reflect their underlying positivity) and –1 for sarcasm (to correct for overstated positivity). The modified classification module, which integrates these adjustments into the R-SST predictions, is referred to as R-HBSC (HumbleBragging and SarCasm).

Table 7 shows that adding this pragmatic layer significantly improves sentiment classification across all metrics. This demonstrates that accounting for implicit cues like humblebragging and sarcasm better aligns model predictions with intended sentiment.

8 Conclusion and Future Work

We introduced the task of automatic humblebragging detection, formalized through our proposed 4-tuple definition. We benchmarked various machine learning, deep learning, and large language models on this task, providing a comparative analysis against human performance. We also demonstrate that our 4-tuple definition significantly improves the zero-shot capabilities of all decoder models. Additionally, we released a synthetic dataset, HB-24, generated using GPT-4o, to facilitate further research. Our experiments and analysis reveal that detecting humblebragging is a challenging task, even for humans. This study lays the groundwork for exploring this intricate linguistic phenomenon and its integration into natural language understanding systems.

Future research could aim to enhance models for identifying humblebragging, fostering a deeper comprehension of this distinct communication style. This may include methods for utilizing contextual cues more effectively. Another valuable direction could involve generating humblebrag captions for images. Additionally, machines could be trained to transform direct brags into humblebrags.

Limitations

The inherent subjectivity of humblebragging complicates the creation of universally agreed-upon labels, as even humans often struggle to classify such statements consistently. Additionally, while machine-generated texts are sophisticated and well-structured, they often lack the spontaneity and imperfections typical of human-authored texts. For instance, the model's inability to use certain casual or curse words, as well as elongated words like *sooooo* or *goood*, which are often present in human-written humblebrags. This creates a mismatch when using synthetic datasets like HB-24, which, despite being a valuable resource, may fail to fully capture the linguistic diversity and subtleties of real-world humblebragging, thereby limiting the generalizability of trained models. Moreover, the task itself remains underexplored, with no prior benchmarks or resources, making it difficult to contextualize results within the larger field of natural language understanding.

References

Mark Alfano and Brian Robinson. 2014. Bragging. *Thought: A Journal of Philosophy*, 3(4):263–272.

Francesco Barbieri and Horacio Saggion. 2014. Modelling irony in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 56–64, Gothenburg, Sweden. Association for Computational Linguistics.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada. Association for Computational Linguistics.

Pushpak Bhattacharyya and Aditya Joshi. 2017. Computational sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Copenhagen, Denmark. Association for Computational Linguistics.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. Association for Computational Linguistics.

Feier Chen, Stephanie Q Liu, and Anna S Mattila. 2020. Bragging and humblebragging in online reviews. *Annals of Tourism Research*, 80:102849.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. PoliSe: Reinforcing politeness using user sentiment for customer care response generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6165–6175, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Haoyu Gao, Ting-En Lin, Hangyu Li, Min Yang, Yuchuan Wu, Wentao Ma, Fei Huang, and Yongbin Li. 2024. Self-explanation prompting improves dialogue understanding in large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14567–14578, Torino, Italia. ELRA and ICCL.

Joana Garmendia. 2018. *Irony*. Cambridge University Press.

Raymond W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 115(1):3.

Montgomery Gole, Williams-Paul Nwadiugwu, and Andriy Miranskyy. 2023. On sarcasm detection with openai gpt-based models. *Preprint*, arXiv:2312.04642.

Shai Gretz, Alon Halfon, Ilya Shnayderman, Orith Toledo-Ronen, Artem Spector, Lena Dankin, Yannis Katsis, Ofir Arviv, Yoav Katz, Noam Slonim, and Liat Ein-Dor. 2023. Zero-shot topical text classification with LLMs - an experimental study. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 9647–9676, Singapore. Association for Computational Linguistics.

Yueguo Gu. 1990. Politeness phenomena in modern chinese. *Journal of Pragmatics*, 14(2):237–257. Special Issue on Politeness.

Xiaobo Guo, Neil Potnis, Melody Yu, Nabeel Gillani, and Soroush Vosoughi. 2024. The computational anatomy of humility: Modeling intellectual humility in online public discourse. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5701–5723, Miami, Florida, USA. Association for Computational Linguistics.

Dan Han, Rong Chen, and Fengguang Liu. 2024. Also on humblebragging: An evaluation of self-image in versailles literature. *Journal of Pragmatics*, 227:4–15.

Irazú Hernández-Farías, José-Miguel Benedí, and Paolo Rosso. 2015. Applying basic features from sentiment analysis for automatic irony detection. In *Pattern Recognition and Image Analysis*, pages 337–344, Cham. Springer International Publishing.

Vera Hoorens, Mario Pandelaere, Frans Oldersma, and Constantine Sedikides. 2012. The hubris hypothesis: You can self-enhance, but you'd better not show it. *Journal of personality*, 80(5):1237–1274.

Stacey Ivanko and Penny Pexman. 2003. Context incongruity and irony processing. *Discourse Processes - DISCOURSE PROCESS*, 35:241–279.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Mali Jin, Daniel Preotiuc-Pietro, A. Seza Doğruöz, and Nikolaos Aletras. 2024. Who is bragging more online? a large scale analysis of bragging in social media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17575–17587, Torino, Italia. ELRA and ICCL.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J. Carman. 2017a. Automatic sarcasm detection: A survey. *ACM Comput. Surv.*, 50(5).

Aditya Joshi, Pranav Goel, Pushpak Bhattacharyya, and Mark Carman. 2017b. Automatic identification of sarcasm target: An introductory approach. *Preprint*, arXiv:1610.07091.

Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 757–762, Beijing, China. Association for Computational Linguistics.

Mikhail Khodak, Nikunj Saunshi, and Kiran Vodrahalli. 2018. A large self-annotated corpus for sarcasm. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Moshe Koppel and Jonathan Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22:100–109.

Dmitry Lamanov, Pavel Burnyshev, Ekaterina Artemova, Valentin Malykh, Andrey Bout, and Irina Piontkovskaya. 2022. Template-based approach to zero-shot intent recognition. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 15–28, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Geoffrey Leech. 1983. *Principles of pragmatics*. Longman, London.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Yanling Lin and Xinren Chen. 2022. Also on humblebragging: Why many chinese posters brag by complaining. *Journal of Pragmatics*, 201:149–159.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

Mufan Luo and Jeffrey T. Hancock. 2020. Humblebragging, self-presentation, and perceptions of (in)sincerity: Chapter 12. Modified self-praise in social media. In María Elena Placencia and Zohreh R. Eslami, editors, *Complimenting Behavior and (Self-)Praise across Social Media: New contexts and new insights*, Pragmatics & Beyond New Series, pages 289–310. John Benjamins Publishing Company.

Abraham Maslow and Karen J Lewis. 1987. Maslow's hierarchy of needs. *Salenger Incorporated*, 14(17):987–990.

Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. 2024. ILLUMINER: Instruction-tuned large language models as few-shot intent classifier and slot filler. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8639–8651, Torino, Italia. ELRA and ICCL.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2021. BERT's the word : Sarcasm target detection using BERT. In *Proceedings of the 19th Annual Workshop of the Australasian Language Technology Association*, pages 185–191, Online. Australasian Language Technology Association.

Widya Paramita and Felix Septianto. 2021. The benefits and pitfalls of humblebragging in social media advertising: the moderating role of the celebrity versus influencer. *International Journal of Advertising*, 40(8):1294–1319.

Daniel Preoţiuc-Pietro, Mihaela Gaman, and Nikolaos Aletras. 2019. Automatically identifying complaints in social media. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5019, Florence, Italy. Association for Computational Linguistics.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Nikunj Saunshi, Sadhika Malladi, and Sanjeev Arora. 2021. A mathematical exploration of why language models help solve downstream tasks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ovul Sezer, Francesca Gino, and Michael Norton. 2015. The psychology of humblebragging. *ACR North American Advances*.

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):52–74.

Abhishek Singh, Eduardo Blanco, and Wei Jin. 2019. Incorporating emoji descriptions improves tweet classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2096–2101, Minneapolis, Minnesota. Association for Computational Linguistics.

Apoorva Singh, Raghav Jain, and Sriparna Saha. 2023. Reimagining complaint analysis: Adopting Seq2Path for a generative text-to-text framework. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Nusa Dua, Bali. Association for Computational Linguistics.

Nancy E Snow. 1995. Humility. *J. Value Inquiry*, 29:203.

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.

Anirudh Srinivasan and Eunsol Choi. 2022. TyDiP: A dataset for politeness classification in nine typologically diverse languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5723–5738, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, and et al. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Hitankshi Trivedi and Vijayalaya Srinivas. 2019. When bragging, be modest: The art of humblebragging. *The International Journal of Indian Psychology, Volume 7, Issue 1, Version 2*, 1:276.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Marek A Vranka, Adéla Becková, and Petr Houdek. 2017. The effects of presenting strengths as weaknesses: Is humblebragging an effective impression management tactic in a job interview?

Zhiyuan Wen, Rui Wang, Shiwei Chen, Qianlong Wang, Keyang Ding, Bin Liang, and Ruifeng Xu. 2023. Rdvi: A retrieval–detection framework for verbal irony detection. *Electronics*, 12(12).

H. Wittels. 2012. *Humblebrag: The Art of False Modesty*. Grand Central Publishing.

Huili Yan, Yuzhi Wei, Chenxin Shen, and Hao Xiong. 2024. Bragging or humblebragging? the impact of travel bragging on viewer behavior. *Tourism Review*.

An Yang, Baosong Yang, and Binyuan Hui et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Qingcheng Zeng and An-Ran Li. 2022. A survey in automatic irony processing: Linguistic, cognitive, and multi-X perspectives. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 824–836, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. 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. *Preprint*, arXiv:2306.05685.

Baiyao Zuo. 2023. "i don't mean to humblebrag"—on the reception of humblebrags from a cognitive-pragmatic perspective. *Journal of Pragmatics*, 218:165–179.

A On the Prevalence of Humblebragging

Although less common compared to sarcasm or irony, humblebragging is not nonexistent, as evidenced by the Google Trends graph reproduced in Figure 4.

The graph clearly shows that the popularity of the term has been increasing gradually over time.

Moreover Reddit has a dedicated r/humblebrag subreddit with 188 thousand members. In contrast r/sarcasm has only 39 thousand members while r/irony has only 60 thousand.

In case of Twitter/X, tweetbinder.com returns the following tweet counts over the past week (search performed on 14 Feb 2025):

- #humblebrag: 72
- #sarcasm: 200
- #irony: 200

Note here that 200 is the limit for the free tier searches.

All of the above statistics correspond to the explicit mentions of the phenomena in question. We suspect implicit humblebragging is much more common in social media. But providing statistics about it is all the more difficult.

Additionally, as observed by Wittels in his book (Wittels, 2012), the phenomena of humblebragging is not recent. In fact only the term for the phenomena is recent. People have been engaging in humblebragging from historical times knowingly or unknowingly.

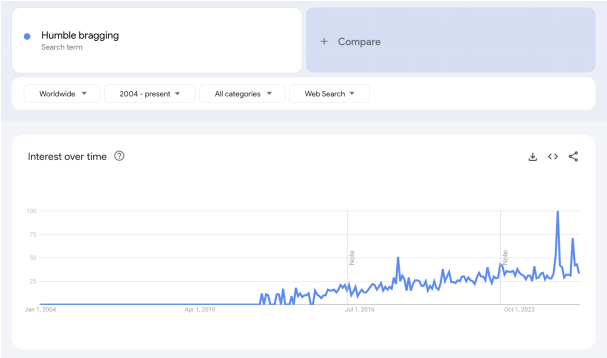


Figure 4: Polpularity of humblebragging.

B How Humblebragging Differs from Irony, Sarcasm, Bragging and Complaint

While irony, sarcasm, and humblebragging rely on indirect communication, bragging and complaints are more direct. Irony presents a contrast between expectation and reality, often meaning the opposite of what is stated. Sarcasm builds on irony but with a sharper, mocking tone, where the surface meaning appears positive, but the intent is negative. Humblebragging disguises self-promotion within a complaint or self-deprecating remark, appearing negative while aiming to impress.

In contrast, bragging and complaints do not rely on hidden meanings. Bragging openly expresses pride, maintaining a positive tone in both surface meaning and intent. Complaints directly convey dissatisfaction, with both their surface and intended polarity being negative. These distinctions are summarized in Table 8.

C Humblebrag Categories

The mapping between Wittel's humblebrag themes to Sezer's humblebrag categories is shown in Table 9.

D Data Generation Prompts

General Prompt

```
You are now a person about to humblebrag
about your recent achievement to attract
people's attention and make them praise
you.But you can't state the obvious. You
have to present it in such a way that it
sounds like a complaint without reducing
the importance of the achievement.
There should be a strong incongruence.
Make sure these are tweets, and keep the
tone casual. Be specific about your
achievements and use diverse topics.
Do not use topics already generated,
and do not follow a pattern for
beginning the text.
```

Prompt with Themes

```
Here is the definition of humblebragging:
a specific type of brag that masks the
boasting part of a statement in a
faux-humble guise. The false humility
allows the offender to boast about their
"achievements" without any sense of shame
or guilt. Humblebrags are usually
self-deprecating in nature.

Now, you are a person who is about to
humblebrag on Twitter with the theme
<theme> and it should sound casual.
Use the above definition and generate
humblebrags.
```

E Dataset Quality Assurance

After synthetic data generation, a manual verification was performed by the first two authors of the paper. In this manual verification step, the main aim was to ensure the selection of high-quality samples for the dataset by:

- Removal of near duplicates.
- Inclusion of diverse categories of humblebrags that represent real-world scenarios.

Moreover, the 4-tuple definition of humblebragging guided the entire manual filtering stage. For instance, out of the following samples, only one was selected, and the rest were discarded:

- "Why can't they just serve normal snacks in first class? Caviar and champagne get so repetitive."
- "Why do they always offer turn-down service on long-haul flights? Sometimes I just want to make my own bed."

Phenomenon	Surface Polarity	Intended Polarity
Irony	Positive or Negative	Opposite of surface polarity
Sarcasm	Positive	Negative
Humblebragging	Negative	Positive
Bragging	Positive	Positive
Complaint	Negative	Negative

Table 8: Surface and Intended Polarity of Different Phenomena

- "Why does the in-flight chef always insist on preparing gourmet meals? Sometimes I just crave a simple sandwich."
- "Why do first-class cabins have private suites? I kind of miss the open seating vibe of economy."

This manual filtering stage was followed by a discussion round wherein the two authors discussed both the filtered-in and filtered-out samples. For cases of disagreement, another round of filtering followed by discussion was performed.

F Binary Classification Framework for Humblebragging Detection

Text Encoding Generation The input text x is converted into a numerical representation $\mathbf{e}$ using any of the available encoding techniques. Formally:

$$\mathbf{e} = f_{\text{encoder}}(x) \tag{2}$$

where:

- x: Input text.
- $\mathbf{e}$: Encoded representation of the text, typically a fixed-dimensional vector.
- f_{encoder}: The encoding function, such as TF-IDF, BERT or a similar pre-trained transformer.

This encoded representation $\mathbf{e}$ captures semantic and contextual information from the input text, enabling effective classification.

Binary Classification Using the encoded representation $\mathbf{e}$, the model predicts the probability $\hat{y} \in [0, 1]$ for the text belonging to the class C_{HB}. The true label y is $y = 1$ for humblebragging and $y = 0$ otherwise.

The Binary Cross-Entropy (BCE) loss for this task is defined as:

$$\mathcal{L}_{\text{BCE}}(y, \hat{y}) = -\frac{1}{N} \sum_{i=1}^{N} \Big(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \Big) \tag{3}$$

where:

- N: Total number of samples in the dataset.
- y_i: True label for the i-th sample ($y_i \in \{0, 1\}$), where $y_i = 1$ indicates a humblebrag.
- $\hat{y}_i$: Predicted probability that the i-th sample is a humblebrag ($\hat{y}_i \in [0, 1]$).

Objective The model is trained to minimize $\mathcal{L}_{\text{BCE}}$ over the dataset, improving its ability to accurately classify texts as humblebragging or non-humblebragging.

G Inference Prompts

User Prompt

```
### Question: Is this a humble brag?
Answer in yes or no only.
### Statement: {data_point['text']}
### Answer:
```

System Prompt

```
A humble brag comprises the following
components:

1. Brag:
   - The segment of the text that
   explicitly conveys the act of
   bragging.

2. Brag Theme:
   - The overarching theme or specific
   category of the brag embedded within
   the statement.
   - Possible categories include:
       - Looks and Attractiveness
       - Achievements
       - Performance at Work
       - Money and Wealth
       - Intelligence
       - Personality
       - Social Life
       - Miscellaneous
```

Category (Sezer et al., 2018)	**Theme** (Wittels, 2012)
Looks and Attractiveness	Ugh, Being Hot Sure Can Be Annoying! Ugh, It's Tough Being a Model Ugh, I'm Too Skinny! Ugh, People Keep Hitting on Me!
Achievements	Ugh, Can You Believe They Included Me on This List? Ugh, I Can't Believe I Won an Award
Performance at Work	Ugh, I'm So Successful
Money and Wealth	Ugh, I Hate Having All This Money!
Intelligence	Ugh, I'm a Genius
Personality	Ugh, I'm So Humble! Ugh, It's Hard Being So Charitable!
Social Life	Ugh, It's So Weird Getting Recognized! Ugh, I Hate People Wanting My Picture and Autograph All the Time Ugh, I'm at an Exclusive Event! Ugh, Being an Author Is Hard! Ugh, How'd I Get Here??? How Is This My Life??? Ugh, I Travel Too Much!
Miscellaneous	Ugh, I Can't Believe I Was Mentioned in This Thing!

Table 9: Mapping Wittel's humblebrag themes to Sezer et al.'s humblebrag categories.

```
3. Humble Mask:
   - The element of the text that adopts
   a modest or self-deprecating tone to
   obscure or mitigate the act of
   bragging.

4. Mask Type:
   - Specifies whether the humble mask
   adopts a modest tone or a
   complaining approach.

Now you are about to classify if a given
sentence is a humble brag or not using
the above definition.
```

H Detailed Experimental Setup

Below we outline the hyperparameter settings used for various models along with a brief discussion of the human annotation.

H.1 Machine Learning Classifiers

We conducted a grid search on various hyperparameters to identify the best combination for each classifier. For the Logistic Regression model, we set the maximum number of iterations (`max_iter`) to 100 to ensure convergence, the regularization strength (`C`) to 0.1 to control overfitting, and a fixed random state (`random_state=42`) for reproducibility. For the Support Vector Classifier (SVC), we used a radial basis function (RBF) kernel (`kernel='rbf'`) to capture non-linear relationships in the data and set the regularization parameter (`C`) to 10.

Unlike other types of textual data, tweets are often informal and unique in their composition. They frequently include *emojis* and *emoticons*, which add emotional or contextual cues. Additionally, tweets commonly feature elongated words (e.g., *soooo* or *goood*) and repeated characters for emphasis or emotional expression. While modern tokenizers utilized by pre-trained networks are designed to handle these emojis and elongated words effectively, traditional machine learning algorithms often struggle with such unconventional text patterns.

Thus, for machine learning classifiers, we incorporated existing pre-processing techniques in

two distinct phases. In the first phase, we replaced all emojis with their corresponding verbal explanations, as suggested by Singh et al. (2019), to retain the semantic information. In the second phase, we utilized *ekphrasis* (Baziotis et al., 2017), a specialized text pre-processing tool, to handle hashtags, elongated and repeated words, URLs, and numeric information. The output was further pre-processed by removing stop words and punctuation, constructing unigram and bigram tokens, limiting the vocabulary to the top 10,000 tokens by frequency, and requiring each token to appear in at least two training samples. The resulting numerical representations, created using TF-IDF, were used as inputs for machine learning classifiers- Support Vector Machine and Logistic Regression.

H.2 Encoder Models

BERT-Large-Uncased (340M) was trained with a learning rate of 5e-3, batch size of 16, and 4 epochs, while RoBERTa-Large (355M) used a learning rate of 5e-4, batch size of 32, and 5 epochs. Both models shared a maximum sequence length of 128, a weight decay of 0.01, a warmup ratio of 0.1, and gradient clipping at 1.0.

H.3 Decoder Models

We evaluated open-source models from Hugging Face—Qwen2.5-7B-Instruct (Yang et al., 2024), Llama 3.1-8B-Instruct (Dubey et al., 2024), Gemma 1.1 7B (IT) (Team et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Vicuna-7B-v1.5 (Zheng et al., 2023)—as well as GPT-3.5 and GPT-4o via OpenAI's API[3]. Outputs were limited to two tokens, and each model was run five times to record average metrics. Prompt details are in Appendix G.

Fine-tuning was conducted using LoRA (Low-Rank Adaptation) with a scaling factor (`lora_alpha`) of 8, rank (`r`) of 16, and targeted attention modules (`q_proj` and `k_proj`). The dataset was split into 80% training and 20% validation with a random seed of 42. Training employed a cosine learning rate scheduler with a learning rate of 1e-5, weight decay of 0.01, and 10 warmup steps, over 4 epochs with a batch size of 8 and gradient accumulation steps of 4. The maximum sequence length was set to 512 tokens, and the `SFTTrainer` was utilized for efficient fine-tuning.

[3] https://platform.openai.com/docs/overview

All experiments were conducted on NVIDIA A100-SXM4-80GB GPUs, utilizing approximately 300 GPU hours in total.

H.4 Human Performance

To evaluate human performance, the test dataset was labeled by three independent annotators, including two with Masters degrees in Linguistics and Arts, and a final-year Masters student in Computer Applications. All were proficient in English and experienced in professional annotation. An initial meeting with the authors covered annotation guidelines and the scope of the task. The annotators first labeled a small subset of the dataset, and disagreements were discussed to ensure consistency. They then labeled the entire test set as *yes* or *no*, indicating the presence or absence of humblebragging. The annotators were compensated according to University norms.

I Inconsistency in Human Annotation

On analyzing the misclassifications made by the lowest-performing human annotator, we observed that the primary factor was not the 4-tuple definition itself. Instead, cultural differences between the source of the tweets and the annotator's background played a significant role, leading to missed contextual cues.

For instance, the following examples were misclassified by the annotator as they failed to identify celebrity status of people mentiond in the example:

- *Sitting next to Penny Marshall at the Lakers Game. #GEEKINGOUT*
- *Tonight: private dinner/event by Miles Davis Estate—his 85th birthday with his family, musicians, media (Beyond humbled/honored to be invited).*
- *Watching myself on Larry King. Achievement diminished in ad break by catheter commercial.*

To address this issue, future studies should consider recruiting annotators from diverse cultural and geographical backgrounds or implementing mechanisms to provide additional contextual information. However, such interventions are beyond the scope of the present study and are proposed as directions for future research.

J Random Gibberish Prompt

```
A Quantum Pancake involves the following
components:

    1. Flapjack Fluctuation:
       - Analyze the positive or negative
       curvature of the syrupy timeline
       as the pancake flips through
       space-time.

    2. Stack Dynamics:
       - Identify the structural
       integrity of the pancake layers
       and their inter-dimensional
       fluffiness coefficient.

    3. Butter Singularities:
       - Highlight the concentrated
       points of creamy chaos where the
       butter both exists and does not
       exist simultaneously.

    4. Maple Entanglement:
       - Describe the sticky phenomena
       where the syrup defies Newtonian
       logic to connect pancakes across
       parallel brunch universes.

Now you are about to classify if a given
sentence is a humble brag or not using
the above definition.
```

K Text Book Definition

```
A specific type of brag that masks
the boasting part of a statement
in a faux-humble guise. The false
humility allows the offender to
boast about their achievements
without any sense of shame or guilt.
Humblebrags are usually
self-deprecating in nature.
```

L Confusion Matrices

The confusion matrices for Qwen, Gemma, Mistral, Vicuna, BERT and RoBERTa are in Figure 5, Figure 6, Figure 7, Figure 8 and Figure 9 (BERT and RoBERTa) respectively.

M Performance Degradation Post Fine-Tuning

We did a full fine-tuning on Vicuna but this had no effect on its performance. We suspect that Vicuna, being an older generation model, is not able to cope up with the newer models. This is also evident through Vicuna's position towards the end of the leaderboards in both Chatbot Arena (https://lmarena.ai/?leaderboard) as well as OpenLLM (https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/).

We tried full fine-tuning on Mistral but due to resource constraints we could not explore it much. Our conjecture is that due to the small size of our dataset and the nuanced nature of our task, Mistral could have experienced catastrophic forgetting. This type of behavior is also reported by Li et al. (2024).

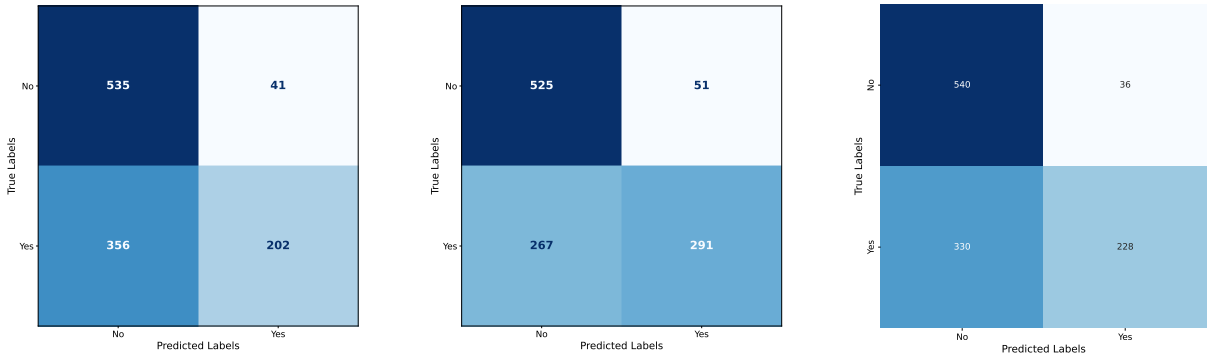
For mitigation, future research directions could be increasing the dataset size, augmenting synthetic data with human-written data and exploring hyperparameter tuning.

N Sentiment-Opposition Model Prompt

```
The Sentiment Opposition Model (SOM) for
Humblebragging consists of the following
components:

    1. Surface Sentiment (SS):
       - The apparent emotional tone of
       the statement, usually negative
       (complaint)
       or neutral (modest).

    2. Intended Sentiment (IS):
       - The actual meaning the speaker
       conveys, which is typically
       positive and self-promotional.

    3. Sentiment Opposition (SO):
       - The contrast between SS and IS.
       If SS is negative/neutral but IS
       implies positive tone,
       opposition exists.

    4. Humblebrag Classification:
       - If SO exists, classify the
       statement as a humblebrag.
       - If SO doesn't exist, classify
       as a non-humblebrag

Now you are about to classify if a given
sentence is a humblebrag or not using
the above definition.
```

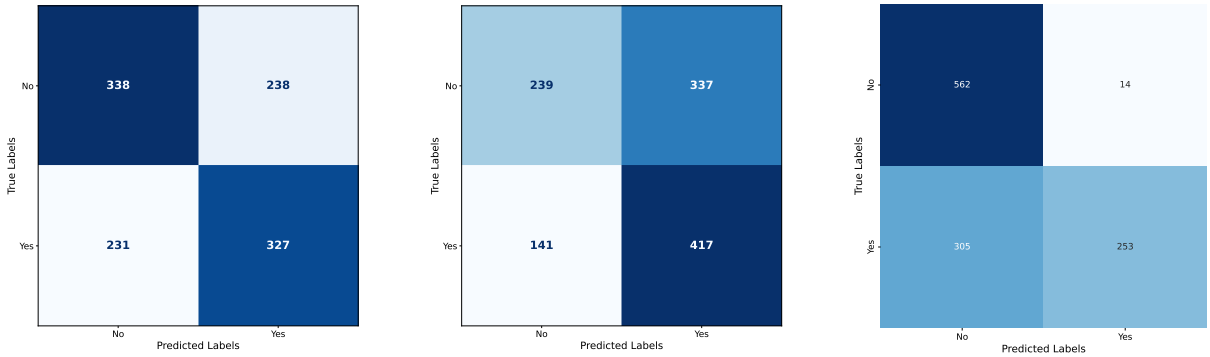
O Humblebragging Component Identification Results

See Table 10 for component identification performance of three different models.



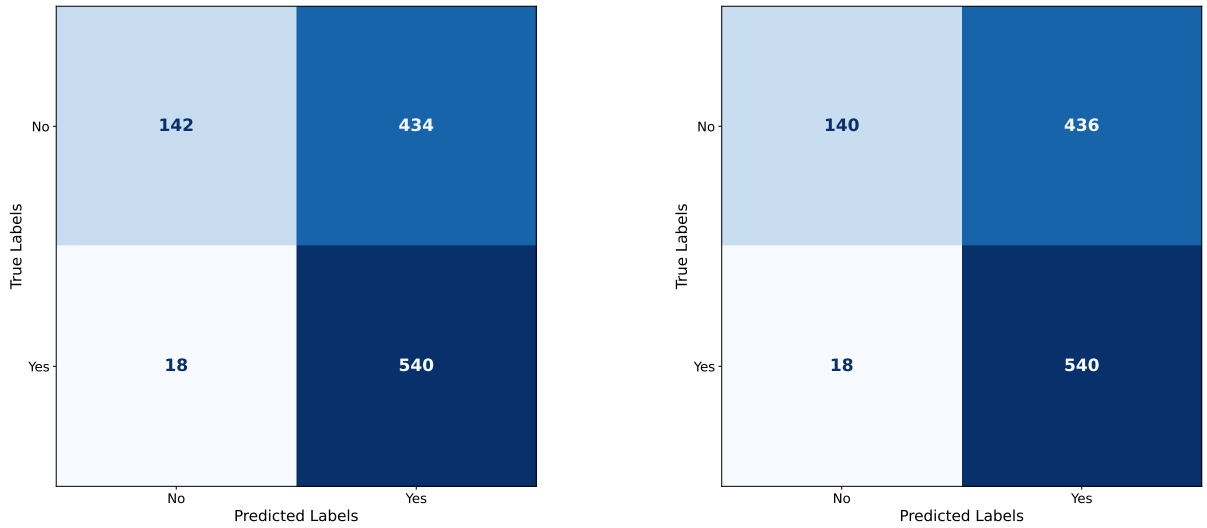
(a) Qwen (Z) (b) Qwen (Z+D) (c) Qwen (F)

Figure 5: Confusion matrices for Qwen Z vs Z+D vs F.



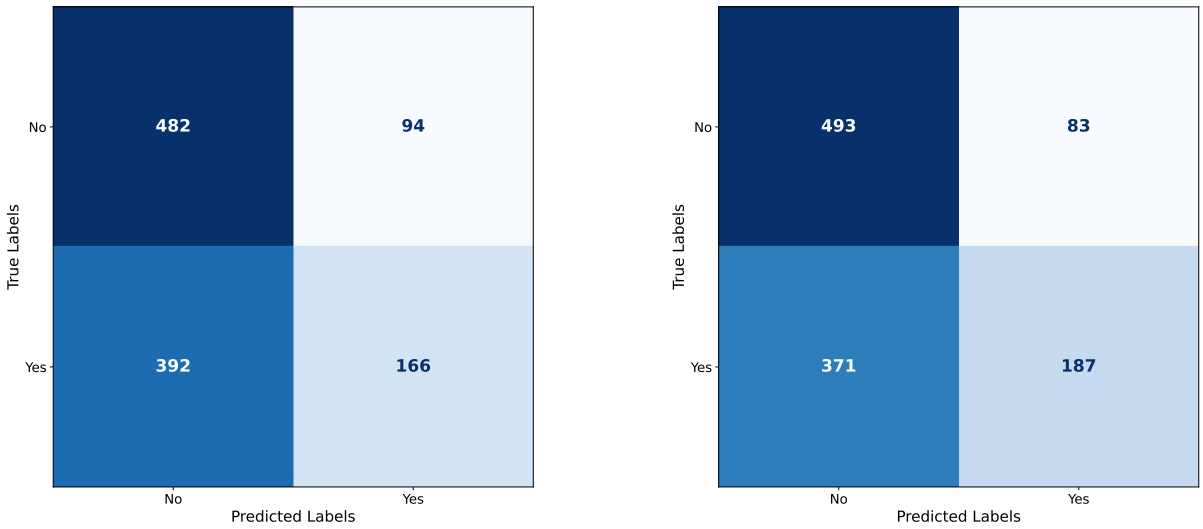
(a) Gemma (Z) (b) Gemma (Z+D) (c) Gemma (F)

Figure 6: Confusion matrices for Gemma Z vs Z+D vs F.



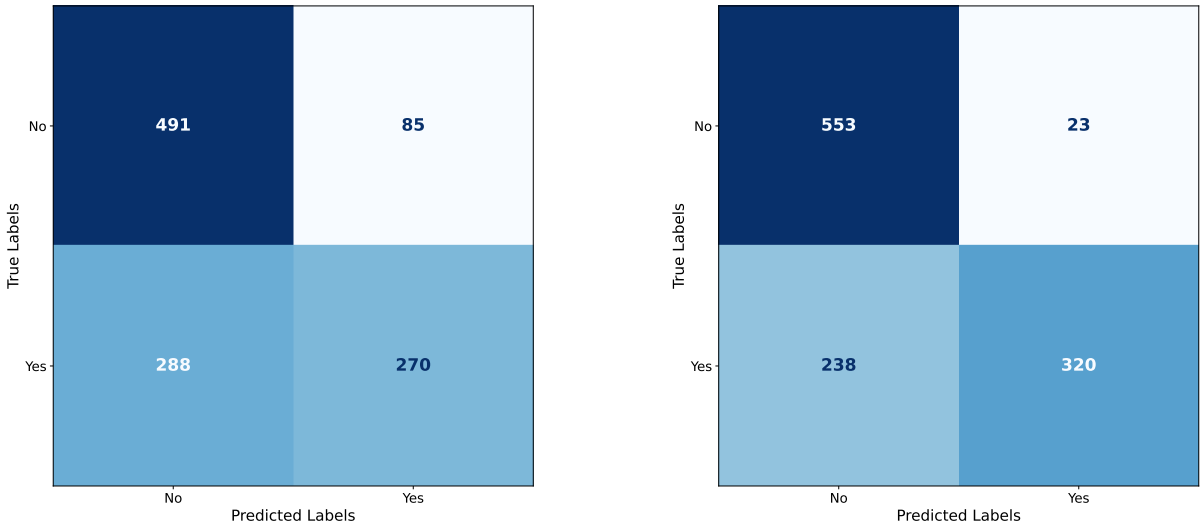
(a) Mistral (Z) (b) Mistral (Z+D)

Figure 7: Confusion matrices for Mistral Z vs Z+D.



(a) Vicuna (Z) (b) Vicuna (Z+D)

Figure 8: Confusion matrices for Vicuna Z vs Z+D.



(a) BERT (F) (b) RoBERTa (F)

Figure 9: Confusion matrices for BERT (F) and RoBERTa (F).

Example 1:	
Input (Gold):	I can't believe they'd give an idiot like me a PhD lol (Brag Theme: Achievements; Mask Type: Complaint)
GPT-4o:	I can't believe they'd give an idiot like me a PhD lol (Brag Theme: Achievement; Mask Type: Self-Deprecating)
Llama-3.1-8B:	I can't believe they'd give an idiot like me a PhD (Brag Theme: Achievements; Mask Type: Modesty)
Mistral-7B:	I can't believe they'd give an idiot like me a PhD (Brag Theme: Intelligence; Mask Type: Complaining)
Example 2:	
Input (Gold):	For the 3rd time in 3 years I've been asked to speak at Harvard, but I've yet to speak at my alma mater. What's a girl gotta do @MarquetteU? (Brag Theme: Achievements; Mask Type: Complaint)
GPT-4o:	For the 3rd time in 3 years I've been asked to speak at Harvard, but I've yet to speak at my alma mater. What's a girl gotta do @MarquetteU? (Brag Theme: Achievements / Intelligence; Mask Type: Complaining)
Llama-3.1:	For the 3rd time in 3 years I've been asked to speak at Harvard, but I've yet to speak at my alma mater.What's a girl gotta do @MarquetteU? (Brag Theme: Achievements; Mask Type: Complaining)
Mistral-7B:	For the 3rd time in 3 years I've been asked to speak at Harvard, I've yet to speak at my alma mater. What's a girl gotta do @MarquetteU? (Brag Theme: Achievement; Mask Type: Complaining)
Example 3:	
Input (Gold):	Will Twitter be available for me in Paris, Milan, or the Maldives? I hope so bc it won't in Hong Kong or Singapore. (Brag Theme: Social Life; Mask Type: Complaint)
GPT-4o:	Will Twitter be available for me in Paris, Milan, or the Maldives? I hope so bc it won't in Hong Kong or Singapore. (Brag Theme: Social Life / Wealth; Mask Type: Complaining)
Llama-3.1:	Will Twitter be available for me in Paris, Milan, or the Maldives? I hope so bc it won't in Hong Kong or Singapore. (Brag Theme: Social Life; Mask Type: Modest)
Mistral-7B:	Will Twitter be available for me in Paris, Milan, or the Maldives? I hope so bc it won't in Hong Kong or Singapore. (Brag Theme: Technology and Connectivity; Mask Type: Complaining)

Table 10: Humblebragging component identification: Model responses with identified brag (red), identified mask (blue), brag theme and mask type.