Making Sense of "Cozy Chaos": Humans and LLMs Differences in Linguistic Creativity

Anca Dinu[1*], **Andra-Maria Florescu**[1*], **Alina Resceanu**[2*]
[1]University of Bucharest, Bucharest, Romania
[2]University of Craiova, Craiova, Romania
anca.dinu@lls.unibuc.ro, andra-maria.florescu@s.unibuc.ro, aresceanu@gmail.com
[*]Equal contribution

Abstract

LLMs have shown remarkable creative capabilities, often outperforming humans in tasks requiring originality and linguistic novelty. However, their creativity differs in nature, raising questions about how humans and machines use language in an original way. This study explores the linguistic dimensions of human and LLM's creativity by analyzing a dataset introduced in previous research. We apply LIWC analysis to identify the proportions of functional words, social and emotional features, and compute the language style matching score to measure stylistic similarity. We also assess readability through standard metrics, perform sentiment analysis, clustering, and binary classification to evaluate how separable the two groups are. The results show that LLMs produce more positive, formal, fluent, and lexically complex responses, whereas human responses are more negative, spontaneous, and socially and experientially grounded.

Introduction

A key motivation for this study is the lack of dedicated evaluation frameworks for the linguistic creativity of large language models (LLMs). Linguistic creativity tasks, such as inventing new words or expressions, require models to go beyond learned patterns and generate novel language forms for which there are no explicit training examples. This makes linguistic creativity tests particularly relevant for assessing computational creativity (CC), as they test the ability of LLMs to creatively manipulate and extend language. Another motivation relevant for CC is to understand the nature of differences in linguistic creativity between humans and machines.

In previous work, (Dinu, Florescu, and Resceanu 2025) introduced a language creativity test in which 24 LLMs and 24 humans were automatically evaluated, showing that the models slightly outperformed humans in most tasks and criteria. However, while performance differences have been documented, the linguistic mechanisms underlying these differences remain underexplored. This study aims to fill this gap: instead of focusing only on performance, we examine how linguistic style, structure, and emotion shape creative expression across both groups. While previous studies have compared human and machine-generated texts mainly through semantic similarity or content-based measures (e.g., (McCoy et al. 2023)), we focus on directly interpretable linguistic and psycho-social features (Gehrmann, Clark, and Sellam 2022); (Zanotto and Aroyehun 2024); (Durward and Thomson 2024). Specifically, we applied the Linguistic Inquiry and Word Count (LIWC) tool (Boyd et al. 2022) to capture grammatical, psycho-social, and emotional features and to compute Language Style Matching (LSM). We also computed readability metrics and performed sentiment analysis, clustering, and classification.

Related Work

The study of CC has flourished in recent years, with the rise and development of LLMs (Chkirbene et al. 2024), their generative potential reaching both their common users and the creative industries (Raza et al. 2025). A detailed analysis of LLMs' creativity was carried out in (Jiang et al. 2024), including a wide range of types of creativity, such as ideational creativity expressed verbally or figuratively, personality creativity, or image generation. A comprehensive review on general artificial creativity is (Ismayilzada et al. 2024).

Most studies explore just the ideational or visual capabilities of LLMs, e.g. (Zhao et al. 2025), which revealed that LLMs are good at elaboration, adding detail and refinement, but weaker in originality or producing genuinely new ideas. However, fewer studies focus on LLMs' creative linguistic capacities, e.g. (Dinu and Florescu 2025; Dinu, Florescu, and Resceanu 2025), which showed a slightly better performance of the LLMs. However, the specific linguistic differences between human and machine answers remain underexplored.

Recent work has also focused on how LLMs handle deeper linguistic and social nuances. For instance, (Weller-Di Marco and Fraser 2024) analyzes how LLMs process morphologically complex words and compound words, showing that while models can generate well-formed word structures, they often rely on surface-level patterns rather than true morphological understanding. Similarly, (Perez Almendros and Camacho-Collados 2024) examined whether LLMs can recognize socially oriented communication patterns, such as *mansplaining*, finding that machines can capture some social cues, but still have difficulties identifying subtle relational and pragmatic contexts. Automatic systems that creatively alter language expressions are proposed in (Gatti, Stock, and Strapparava 2021) with practical goals, such as grabbing the reader's attention or aiding in concept recall. Noting thematic distinctions between the

	function	i	ipron	det	number	prep	conj	negate	verb
MeanLLMs	525.58	0.00	0.00	103.83	11.58	271.83	141.67	4.33	177.13
MeanHumans	814.50	33.42	22.71	255.17	41.17	485.13	46.58	36.33	317.21

Table 1: Comparison of human and LLM mean scores across function word categories.

	affiliation	allnone	cause	discrep	differ	space	auditory
MeanLLMs	82.25	4.33	30.54	8.96	3.58	191.42	292.38
MeanHumans	150.79	20.96	8.67	81.08	33.92	266.00	146.33

Table 2: Comparison of human and LLM mean scores across structural features.

human and AI-generated literature, (Durward and Thomson 2024) examined the language used in creative works.

Our study complements this previous research by analyzing the linguistic and stylistic features of the creative output of humans and machines.

Dataset and Experimental setup

We use the dataset introduced by (Dinu, Florescu, and Resceanu 2025) that comprises the responses of 24 humans and 24 LLMs to a language creativity test, totaling 2304 responses. The eight language creativity tasks are given in the Appendix. The human participants were volunteer students in Humanities (English major), non-native speakers of English, with an English proficiency level of B2 and above, aged between 19 and 25, 8 males and 16 females. It can be argued that the test respondents being non-native English speakers is a limitation of the study. Nevertheless, an article that proposes a language creativity test focused on word formation only (Körtvélyessy et al., 2022) states that "there is no principled difference between native speakers and non-native speakers in their ability to form new complex words and interpret/predict the meaning of novel/complex word provided that the non-native speaker has a standard command of a particular language [(...)] and that his/her world knowledge and experiences are comparable to those of common native speakers". Language creativity manifests itself in non-native speakers in relevant ways, as explained in (Zipp, 2019).

We experimented in Colab, with open sourced libraries: pandas[1], NumPy[2], scikit-learn[3], matplotlib[4],adjustText[5], textstat[6], Seaborn[7], and transformers[8]. For coding assistance, we used ChatGPT(4o)[9] and Claude(3.7)[10].

[1]https://pandas.pydata.org/

[2]https://numpy.org/

[3]https://scikit-learn.org/stable/

[4]https://matplotlib.org/

[5]https://adjusttext.readthedocs.io/en/latest/

[6]https://textstat.org/

[7]https://seaborn.pydata.org/

[8]https://pypi.org/project/transformers/

[9]https://chatgpt.com/

[10]https://claude.ai/

LIWC analysis

We extracted linguistic and psicho-social features from the dataset with the LIWC-22 tool (Boyd et al. 2022), to capture how humans and machines use language, their stylistic, emotional, and cognitive patterns. We obtained a low to moderate LSM score between humans and LLMs (0.43), suggesting substantial differences in the way humans and LLMs use language creatively.

Table 1 shows the human and LLM mean scores across function words. Human responses show a richer and more varied use of function words such as pronouns, determiners, prepositions, and verbs. This is evident in phrases like "I am miserable", "would be drama queens" or "the nightmare of corporatists" where pronouns (I, it), prepositions (in, of), determiners (the, a), and verbs (would) create connected, dynamic expressions. Such constructions make human language sound more relational, narrative, and experiential, reflecting that people draw on lived and social contexts when being creative. Humans also used a lot of negations ("not fully developed" for *childish*, "a sea of untold stories", "unhappy people") because they often think and create through contrast, irony, or dissatisfaction. Saying what something is *not* is a quick and expressive way to show tension or attitude, which fits natural and emotional language. LLMs, on the other hand, avoided frequent negations because their outputs are optimized for fluency, balance, and positive tone. Their responses also contained fewer function words and more self-contained noun phrases, like "pocket-sized thunderstorms" (for angry dogs), "customer service gladiators supreme" (for "I wanna speak to the manager" kind of people), or "(a sea of) endless serenity", which show a well-structured style, but one that sounds detached from real experience and less natural.

Table 2 reveals that human responses are more socially (higher *affiliation*), emotionally (higher *allnone*), and cognitively engaged (higher *discrep* and *differ*). These patterns suggest that human creativity is rooted in subjectivity and self-reflection, since people express themselves through opposites, doubt, and social perspective. In contrast, LLM responses show greater *causal* coherence and *auditory* description, producing structured and stylistically balanced answers that sound fluent but lack the relational nature of human creativity. The higher *space* scores for humans further point to their tendency towards spatially grounded im-

	prosocial	moral	socrefs	family	friend	female	male
MeanLLMs	41.67	63.54	203.08	5.63	15.17	9.46	4.75
MeanHumans	23.25	120.25	446.13	54.79	35.50	94.17	34.71

Table 3: Comparison of human and LLM mean scores across social features.

	home	work	money	relig	physical	health	illness	wellness	mental	food	death	fatigue	allure
MeanLLMs	23.88	124.13	37.33	34.71	158.88	70.38	16.00	39.75	0.00	36.50	4.63	4.79	110.96
MeanHumans	46.50	236.92	66.63	111.08	496.75	127.96	31.46	22.83	16.67	115.42	42.54	15.50	236.33

Table 4: Comparison of human and LLM mean scores across personal features.

agery ("storm in a cup", "fury in a fur coat"), while LLMs' high auditory score indicates formulaic figurative speech ("echoes of calm", "whispers of serenity"). Moreover, human language is deeply relational and morally evaluative, whereas LLM language is neutral, polite, and impersonal, as can be seen in table 3. Humans used more moral terms ("guilty", "innocent", "good", "bad") and social references ("we," "they," "my friend"). LLMs, in turn, produced more prosocial but abstract language ("helpful", "friendly", "kindly"), showing polite, controlled positivity.

Table 4 reinforces that human creativity draws on emotional and existential experiences, including body, home, work, health, food, and death, often with irony or emotional depth ("angry mums" for "I wanna talk to the manager kind of people", "mummy's baby" or "parents' epic failure" for *childish*, "cozy office work stress" or "cozy exam taking" for the oxymoron task). LLMs, on the other hand, produced emotionally regulated 'proper' language, avoiding *illness*, *death* or *fatigue* using instead wellness terms ("cozy chaotic serenity", "chaotic Zen garden meditation hour").

Comparing the frequency of motivational language across the categories in table 5 reveals that human responses show stronger desire and effort (e.g. "tryhardish", "win", "fail", "dream", and "fight"), and irony or references to everyday realism (e.g., "manager hunting warriors", "refund addicts", "let the show begin people", "in need of a therapist" for "I wanna talk to the manager kind of people"). LLMs' style is safe and positive, avoiding any display of power or personal ambition and keeping the tone neutral, emotionally balanced, and collective. This is in line with the *tone* score: LLMs - 2286.88 and humans - 1169.42.

The BigWords score, which measures how many long words (six or more letters) appear in a text, indicates that LLMs (score 6354.96) use longer words than humans (4851.58), creating more complex but harder-to-read constructions (e.g. "Busybody Blabbermouth", "Scandal-spreading Scuttlebutt", "Neighborhood hearsay distributor" for *gossiper)*. Humans preferred shorter and simpler words (e.g., "scandal seeker", "story spreader", "muddy tongue") giving their answers a more vivid and spontaneous feel.

Readability

To analyze differences in readability between the two groups, we computed six metrics with the *textstat* library: Flesch Ease, Flesch-Kincaid, Gunning Fog, ARI, Coleman-Liau, and Dale-Chall, represented in Figure 1 from the Appendix. For all metrics, the answers produced by the LLMs are consistently more difficult to read. Humans formed shorter, simpler, and more familiar expressions that sound direct and conversational, whereas LLMs used longer, more vague, and complex words, which makes their answers fluent but heavier to read. This fits with (McCoy et al. 2023; Zanotto and Aroyehun 2024), who showed that LLMs prefer formal and grammatically complete phrasing, whereas humans favor spontaneity and clarity. The almost 0 p-values obtained for the Welch tests confirm that the differences in mean readability between humans and LLMs are statistically significant, as shown in Table 6 from the Appendix.

Sentiment Analysis

To compare the emotional load of humans and LLMs, we performed sentiment analysis for all 48 files, by using a transformer-based approach via Hugging Face's pre-trained BERT model[11]. Each individual file was assigned a percentage of positive, neutral, and negative content. The distribution of individual sentiments with stacked horizontal bar charts listed in figure 2 from the Appendix. In the first 15 most negative individual responses, only 3 are LLMs and 12 humans, whereas, symmetrically, in the first 15 most positive individual answers only 3 are humans and 12 LLMs. This might be due to the LLM having their content filtered to avoid toxicity and negativity and to humans expressing creativity by the use of emotional contrast, irony, humor, or realism (e.g. "not fully matured" or "not quite grown up" for *childish*, a sea of... "never ending troubles", "unsolicited advice from that aunt", "broken hearts" or "tears and despairs").

Visualization

To visualize systematic differences in creativity answers between humans and LLMs for all 48 individual text files, we performed Principal Component Analysis (PCA) on two types of features: word usage patterns extracted with TF-IDF vectorization, which highlight lexical and topical differences, and writing style and structural features, which underline stylistic differences rather than word choice. The first

[11] https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

	need	want	acquire	lack	reward	risk
MeanLLMs	4.79	7.25	0.54	9.92	9.33	13.67
MeanHumans	12.67	40.38	18.00	26.00	37.79	23.21

Table 5: Comparison of human and LLM mean scores across motivational features.

method produced the visualization in figure 3 from the Appendix. The individuals are moderately clustered into two groups, with some overlappings in the densely populated center, and some individuals placed far away from the center. This indicates that word usage is different across the two groups, but a perfect discrimination is not possible based only on this type of feature. For the second method, we used a deep learning-based style embedding model[12]. This method generates contextual embeddings via transformers, capturing nuances of writing style, tone, and structure. This time, the PCA visualization, given in figure 4 from the Appendix, shows that the two classes separated fairly well, almost linearly, with only two humans grouped with the LLMs and with ChatGPT integrated within the human group. This result suggests that stylistic features are good discriminators between the two classes. The two human outliers are participants who responded in a more formal and structured manner, while the proximity of ChatGPT to the human cluster may be caused by its extensive exposure to human interaction during training and refinement, since, unlike other models, ChatGPT has been continuously improved through reinforcement learning from human feedback (RLHF) (Lambert 2025) and widespread everyday use.

Binary Classification

The results of all previous sections indicate that there are relevant features that discriminate between humans and LLMs. Therefore, we experimented with several classical Machine learning (ML) techniques (Suport Vector machine (SVM), Random Forest (RF), Multinomial Naive Bayes, Logistic Regression) and the DistilBERT[13] transformer, to automatically classify the two groups. As training set, we considered all separate responses from all humans and machines (1152 each). The best performance for ML techniques was obtained with Logistic Regression with TF-IDF vectorization at both character and word level, with an accuracy of 0.68. The confusion matrix shown in 5 (left) from the Appendix reveals 66 humans misclassified as LLMs and 80 LLMs misclassified as humans. This modest performance is likely due to the presence of invented or novel words that the TF-IDF representation cannot capture effectively.

The transformer model outperformed the ML techniques, obtaining an accuracy of 0.75, in the third epoch, as we can see in table 7. The confusion matrix shown in 5 (right) from the Appendix reveals a consistent improvement over the ML algorithms, indicating that deep contextual representations better capture subtle linguistic and invented-word patterns.

[12]https://huggingface.co/AnnaWegmann/Style-Embedding

[13]https://huggingface.co/distilbert/distilbert-base-uncased

Conclusions

In this study, we compared the linguistic creativity of humans and LLMs through a detailed computational and psycholinguistic analysis of 2304 creative responses, using LIWC linguistic profiling, readability metrics, sentiment analysis, visualization, and classification methods. The results reveal that LLMs tend to produce language that is more complex, formal, and lexically dense, marked by structural regularity and positive tone. Their responses are balanced and fluent, but display limited emotional range and spontaneity. This pattern points to exploratory creativity: LLMs appear to generate novel combinations within an already learned linguistic and probabilistic space, rather than transforming the underlying generative rules (i.e. attaching the suffix *-ish* to standard bases like adjectives or nouns, such as in "quikish" and "chirpish"). In contrast, human often redefine or break the rules, creating borderline interpretable utterances (i.e. attaching the suffix *-ish* to non-standard bases, like verbs, "loveish", "hatish", "scrollish", or producing semantically deviated examples, "bottleish", "teethable"), indicative of transformational creativity (Boden 1996; Boden 2004).

Additionally, human creativity emerged as more relational, contextually and emotionally grounded, often expressed through negation (i.e. "no lifer", "no brainer" for 'I wanna speak to the manager' kind of people'), irony (especially in word formation tasks, i.e. "Americanish", "parentish", "wifeable") and experientially or socially referenced (i.e. "childish" paraphrased as "Justin Bieber fan"). These findings suggest that linguistic creativity in humans is affectively and socially embedded, while LLM creativity is optimized toward stylistic coherence and politeness.

The visualization and classification results further support these distinctions: stylistic features proved to be more discriminative than lexical ones. Transformer-based classifiers achieved moderate accuracy (0.75), indicating that while LLM and human language styles differ systematically, overlap still exists. ChatGPT appears stylistically closer to humans, possibly due to reinforcement learning from human feedback and conversational refinements.

Our findings show that although LLMs can mimic human creativity, they lack the emotional and experiential depth.

Limitations

The study relies on a single dataset (Dinu, Florescu, and Resceanu 2025), which, although balanced, covers limited tasks and may not reflect the full range of linguistic creativity. In addition, the strategies of the 48 participants may not generalize to broader populations or newer models. Future work should include larger, more diverse datasets, human evaluation, and cross-linguistic comparisons.

References

[Boden 1996] Boden, M. A. 1996. Chapter 9 - creativity. In Boden, M. A., ed., *Artificial Intelligence*, Handbook of Perception and Cognition. San Diego: Academic Press. 267–291.

[Boden 2004] Boden, M. 2004. *The Creative Mind: Myths and Mechanisms*. Routledge.

[Boyd et al. 2022] Boyd, R. L.; Ashokkumar, A.; Seraj, S.; and Pennebaker, J. W. 2022. The development and psychometric properties of liwc-22. *Austin, TX: University of Texas at Austin* 10.

[Chkirbene et al. 2024] Chkirbene, Z.; Hamila, R.; Gouissem, A.; and Devrim, U. 2024. Large language models (llm) in industry: A survey of applications, challenges, and trends. In *2024 IEEE 21st International Conference on Smart Communities: Improving Quality of Life using AI, Robotics and IoT (HONET)*, 229–234.

[Dinu and Florescu 2025] Dinu, A., and Florescu, A.-M. 2025. Testing language creativity of large language models and humans. In Hämäläinen, M.; Öhman, E.; Bizzoni, Y.; Miyagawa, S.; and Alnajjar, K., eds., *Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities*, 426–436. Albuquerque, USA: Association for Computational Linguistics.

[Dinu, Florescu, and Resceanu 2025] Dinu, A.; Florescu, A.-M.; and Resceanu, A. 2025. A comparative approach to assessing linguistic creativity of large language models and humans. *Procedia Computer Science* 270:1292–1301. 29th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2025).

[Durward and Thomson 2024] Durward, M., and Thomson, C. 2024. Evaluating vocabulary usage in LLMs. In Kochmar, E.; Bexte, M.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Tack, A.; Yaneva, V.; and Yuan, Z., eds., *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 266–282. Mexico City, Mexico: Association for Computational Linguistics.

[Gatti, Stock, and Strapparava 2021] Gatti, L.; Stock, O.; and Strapparava, C. 2021. *Cognition and Computational Linguistic Creativity*. Cham: Springer International Publishing. 1–39.

[Gehrmann, Clark, and Sellam 2022] Gehrmann, S.; Clark, E.; and Sellam, T. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text.

[Ismayilzada et al. 2024] Ismayilzada, M.; Paul, D.; Bosselut, A.; and van der Plas, L. 2024. Creativity in ai: Progresses and challenges.

[Jiang et al. 2024] Jiang, X.; Tian, Y.; Hua, F.; Xu, C.; Wang, Y.; and Guo, J. 2024. A survey on large language model hallucination via a creativity perspective.

[Lambert 2025] Lambert, N. 2025. Reinforcement learning from human feedback. arXiv preprint.

[McCoy et al. 2023] McCoy, R. T.; Smolensky, P.; Linzen, T.; Gao, J.; and Celikyilmaz, A. 2023. How much do language models copy from their training data? evaluating linguistic novelty in text generation using RAVEN. *Transactions of the Association for Computational Linguistics* 11:652–670.

[Perez Almendros and Camacho-Collados 2024] Perez Almendros, C., and Camacho-Collados, J. 2024. Do large language models understand mansplaining? well, actually... In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 5235–5246. Torino, Italia: ELRA and ICCL.

[Raza et al. 2025] Raza, M.; Jahangir, Z.; Riaz, M. B.; Saeed, M. J.; and Sattar, M. A. 2025. Industrial applications of large language models. *Scientific Reports* 15(1):13755.

[Weller-Di Marco and Fraser 2024] Weller-Di Marco, M., and Fraser, A. 2024. Analyzing the understanding of morphologically complex words in large language models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1009–1020. Torino, Italia: ELRA and ICCL.

[Zanotto and Aroyehun 2024] Zanotto, S. E., and Aroyehun, S. 2024. Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models.

[Zhao et al. 2025] Zhao, Y.; Zhang, R.; Li, W.; and Li, L. 2025. Assessing and understanding creativity in large language models. *Machine Intelligence Research* 22(3):417–436.

Acknowledgments

Research supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS - UEFISCDI, project SIROLA, number PN-IV-P1-PCE-2023-1701, within PNCDI IV.

Appendix

The eight tasks of the language creativity test:

I. Word/Compound Formation

1. Combine two words to form a new one that describes a specific concept (A. a person completely relying on chatbots, B. a mixed feeling of satisfaction and guilt)
2. Complete a given word with another to form a new compound (A. -ish, B. -able)
3. Continue a certain series of derived words (A....ice, B....summoner)
4. Invent new words that would fit within the same semantic field as a set of words (A. dog, doggy, puppy, canine, bark, paw, woof, sniff, bitch; B. pig, piggy, pork, oink, swine, hog, piglet, snout, boar)

II. Figures of speech

1. Describe familiar ideas by giving novel alternative names (A. angry small dogs; B. the "I wanna speak to the manager" kind of people)
2. Complete blank spaces so as to create an original metaphorical meaning (A. a/some...later; B. a sea of...)
3. Continue the given words with unusual or funny opposites (A. cozy...; B. chaotic...)
4. Give one original harsher (pejorative) expression, one equivalent expression, and one milder (appreciative) expression for a word (A. gossiper; B. childish)

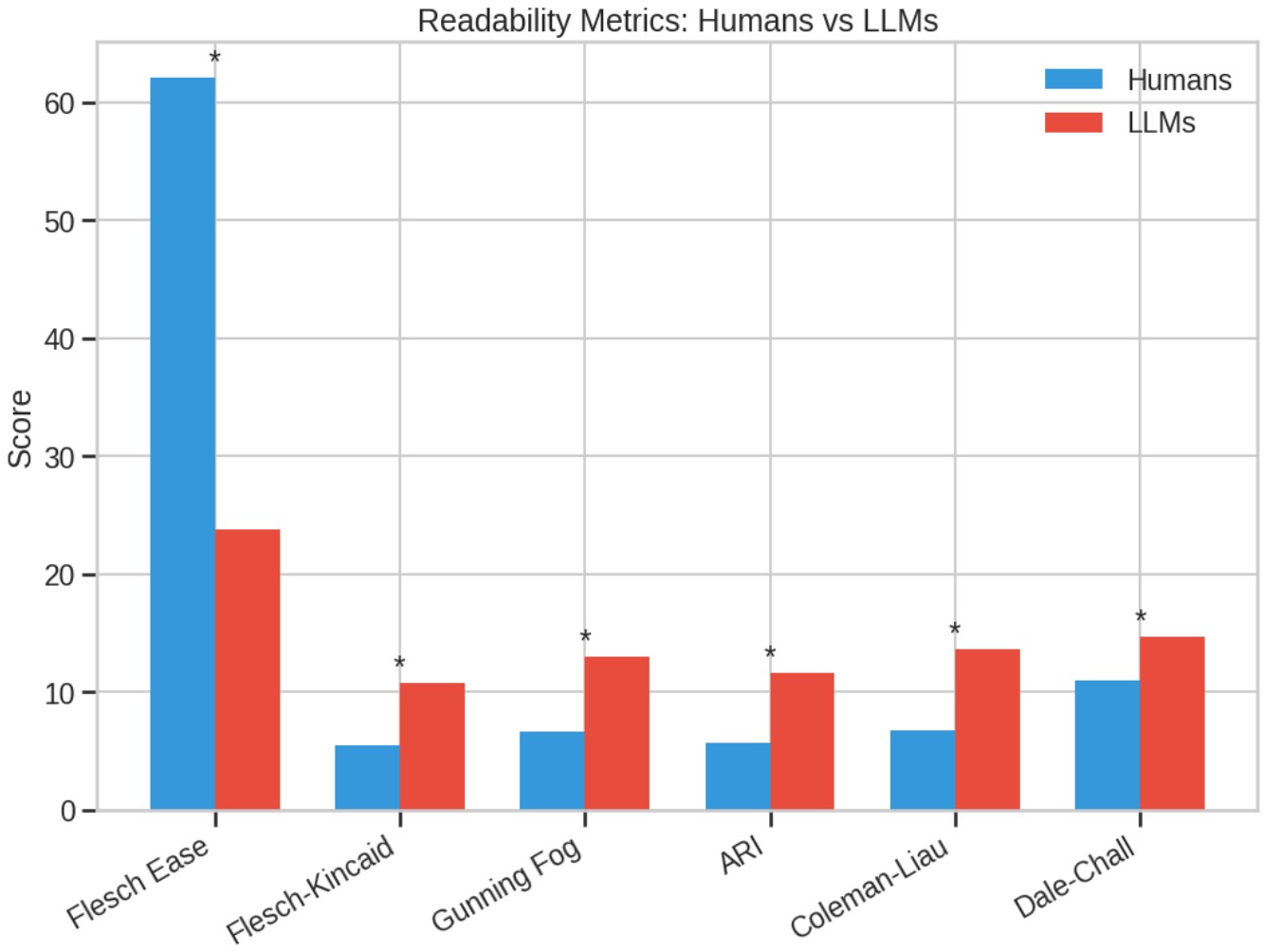

Figure 1: Readability metrics. For the Flesch Ease, the higher the score, the easier the text is to read, as opposed to all other metrics, so the pattern is the same.

Figure 2: Sentiment scores for all 48 individual participants, ordered by their negative sentiment (left), and by their positive sentiment (right).

Metric	Human Mean	LLM Mean	Diff.	p-value
flesch_kincaid_grade	5.48	10.81	5.33	1.95×10^{-17}
flesch_reading_ease	62.07	23.78	-38.29	2.66×10^{-17}
coleman_liau_index	6.69	13.59	6.90	3.00×10^{-17}
automated_readability_index	5.70	11.56	5.86	3.42×10^{-17}
dale_chall_readability_score	11.02	14.72	3.70	1.08×10^{-14}
gunning_fog	6.65	12.98	6.32	7.04×10^{-13}

Table 6: Comparison of readability metrics between human and LLM-generated texts.

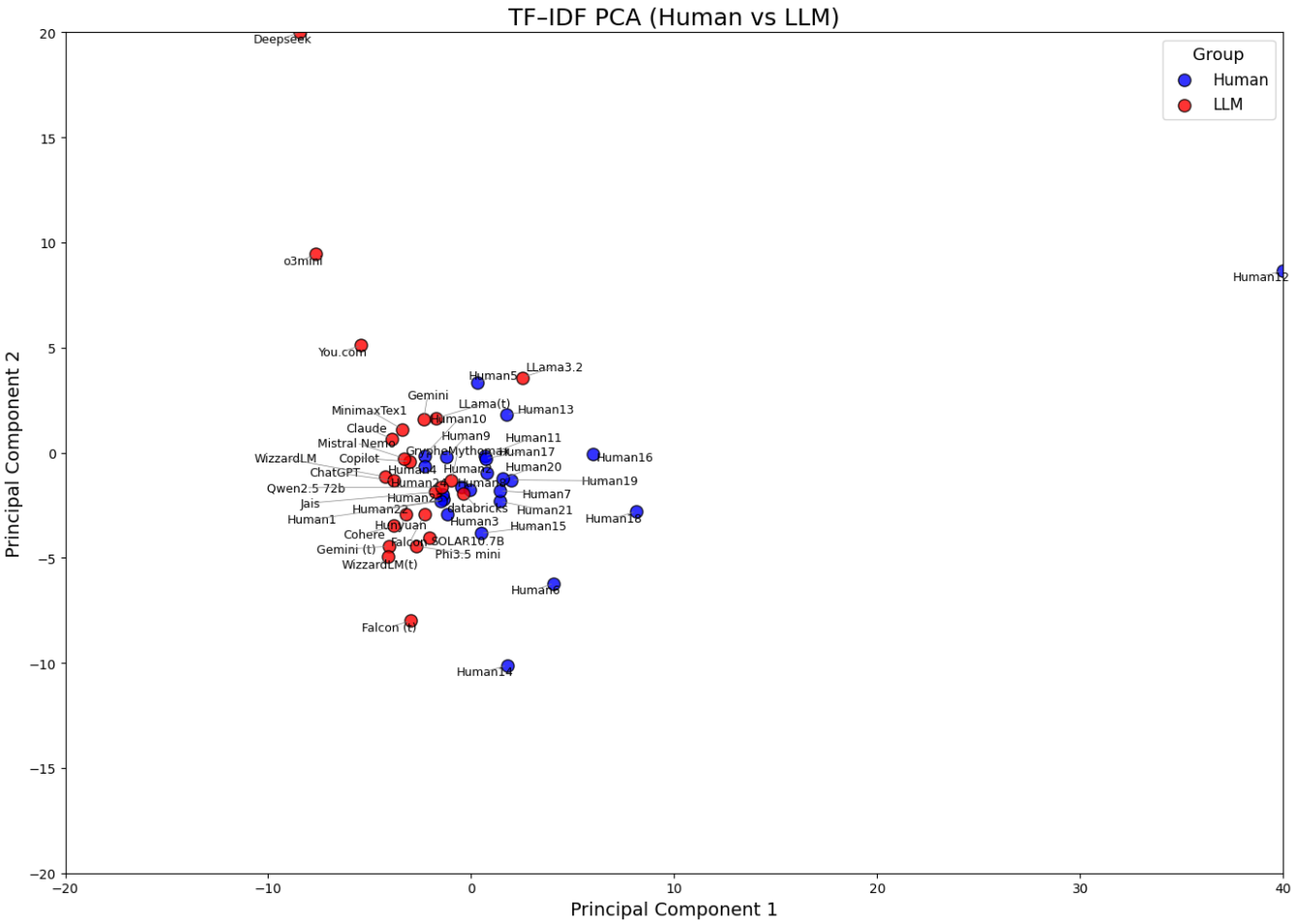

Figure 3: TF-IDF PCA.

	Machine Learning			Deep Learning		
Model	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Humans	0.67	0.71	0.69	0.73	0.80	0.77
LLMs	0.69	0.65	0.67	0.78	0.71	0.74
Accuracy		0.68			0.68	

Table 7: Machine Learning and Deep Learning classification.

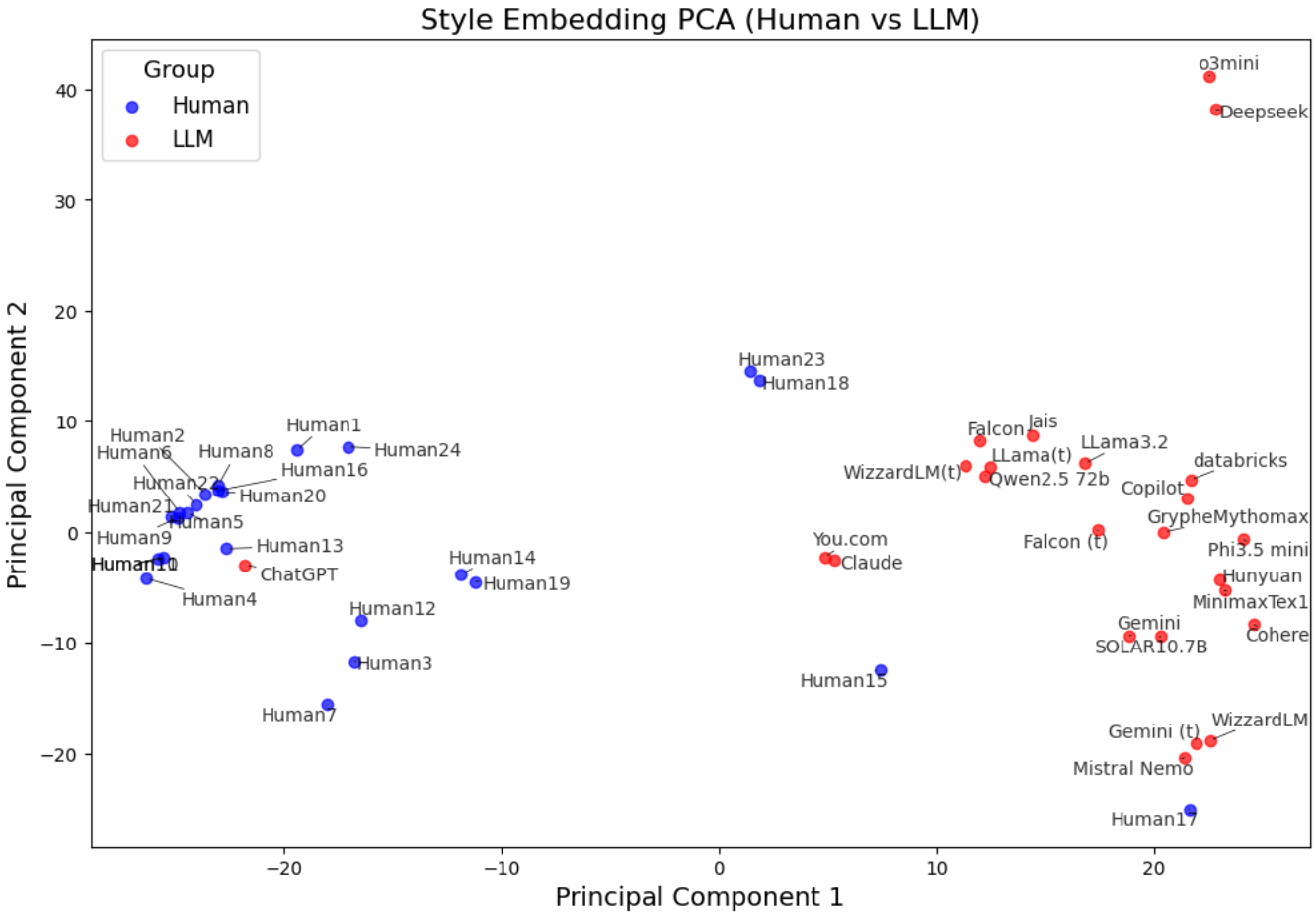

Figure 4: Style embeddings PCA.

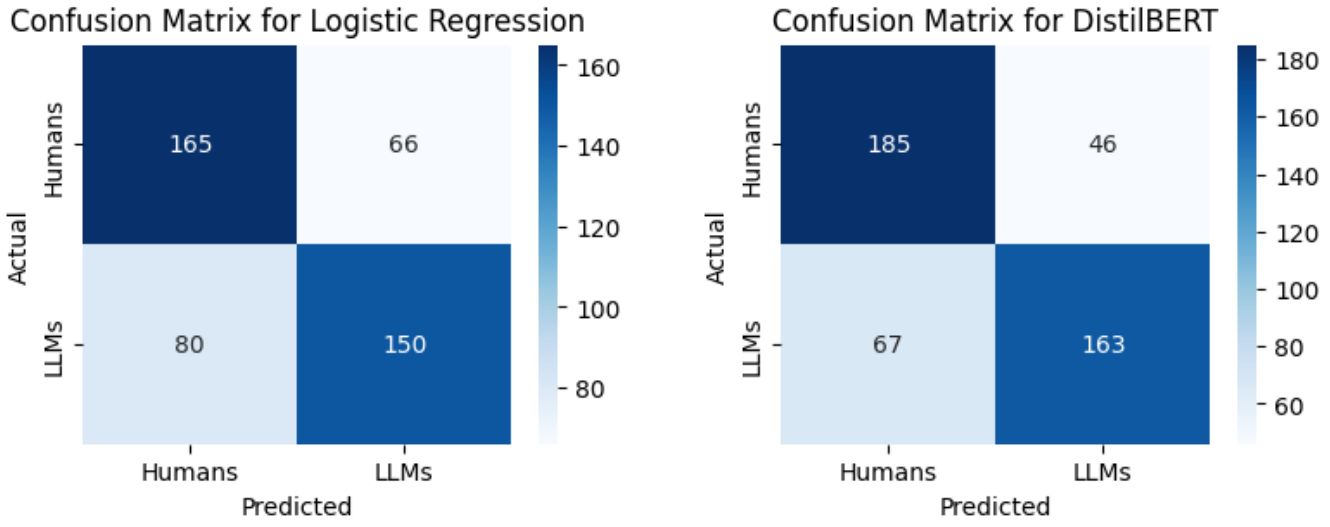

Figure 5: Confusion Matrix for Logistic Regression (left), and for DistilBERT (right).