Human vs LLM Language Comparison

Naama Avni 208523456

1. Research Description

This report presents a comparative analysis between human-written comments on Reddit and responses generated by four Large Language Models (LLMs): Claude-Sonnet, GPT-4.1-mini, Llama, and Qwen. The objective of this study was to identify linguistic "fingerprints" that distinguish human language from machine-generated text. I analyzed the text using 12 distinct metrics covering syntactic complexity, lexical diversity, and structural patterns.

2. Methodology: Metrics Explanation

The following metrics were calculated to quantify the differences:

- **1. Mean Sentence Length:** Average number of words per sentence. Longer averages may indicate run-on sentences or higher complexity.
- **2. Sentences per Comment:** Average number of sentences contained in a single response.
- **3. Type-to-Token Ratio (TTR):** Measure of lexical diversity calculated as Unique Words / Total Words. Heavily biased by text length.
- **4. Standardized TTR:** Length-normalized lexical diversity calculated on a fixed sample (50,000 tokens) to allow fair comparison.
- **5. Root TTR:** Lexical diversity measure using square root normalization: Unique Types / sqrt(Total Tokens).
- **6. MTLD Score:** Measure of Textual Lexical Diversity. Calculates how many words a text maintains before TTR drops below a threshold (0.72). Higher scores indicate richer vocabulary.
- **7. Mean Parse Depth:** Syntactic complexity measure using the depth of the dependency parse tree. Depth 3-5 is simple; 9+ is complex.
- **8. Avg Word Length:** Average character length of words, often acting as a proxy for vocabulary formality.
- **9. Burstiness:** Coefficient of variation of sentence lengths (Std Dev / Mean). Higher values indicate natural human variation; lower values indicate robotic uniformity.
- **10. Punctuation Diversity:** Count of unique punctuation marks used, indicating expressive richness.
- **11. Contraction Rate:** Percentage of words that are contractions (e.g., 'don't', 'it's'). High rates indicate informal/conversational tone.
- **12. Error Score (per 1K words):** Count of mechanical errors (repeated words, spacing issues, capitalization) normalized per 1,000 words.

3. Results

Metric	Human	Claude	GPT4	Llama	Qwen
Mean Sentence Length	16.36	19.07	18.61	34.55	14.92
Sentences per Comment	9.55	8.74	6.8	4.64	9.32
TTR	0.0614	0.0513	0.0646	0.0186	0.0305
Standardized TTR	0.2176	0.2184	0.2204	0.1624	0.1808
Root TTR	81.85	61.36	61.62	29.96	36.28
MTLD Score	129.21	214.94	200.64	93.65	116.89
Mean Parse Depth	4.72	5.33	5.35	7.41	4.74
Avg Word Length	4.45	5.05	4.97	4.73	4.63
Burstiness	0.75	0.44	0.42	0.98	0.49
Punctuation Div	32	31	32	31	32
Contraction Rate	1.67%	3.52%	0.14%	4.07%	1.41%
Error Score	16.8	4.04	2.77	1.93	1.88

4. Impressions and Conclusions

Based on the computed metrics, several distinct patterns emerge distinguishing human writing from LLM generation:

1. The Perfection Paradox (Error Score):

The most striking difference is in the Error Score. Human text is significantly "messier," with a score of 16.80 compared to the models, which range from 1.88 to 4.04. This validates the hypothesis that LLMs are trained to be grammatically hyper-correct. The presence of formatting errors, repeated words, and spacing issues is a strong signature of human typing.

2. Structural Variety (Burstiness):

Humans exhibit a high Burstiness score (0.75), indicating a natural variation between short and long sentences. The closed models (Claude, GPT-4) and Qwen are more robotic/uniform, with scores around 0.42-0.49. Llama is an outlier (0.98), but this might be interpreted alongside its massive Mean Sentence Length (34.55).

3. Lexical Richness (MTLD, Root TTR & Word Length):

Closed-source models (Claude and GPT-4) exhibit significantly higher MTLD scores (~214 and ~200) compared to humans (~129), indicating more consistent lexical diversity throughout their text. However, humans show a remarkably higher Root TTR (81.85 vs Claude 61.36, GPT4

61.62), suggesting humans use a wider overall vocabulary across the full dataset. This reveals distinct strategies: LLMs maintain consistent diversity with formal vocabulary, also reflected in their higher average word lengths (5.05 and 4.97 vs Human 4.45). while humans show broader vocabulary range with simpler words (4.45 average length). The pattern suggests LLMs are tuned for sophisticated consistency, while human writing displays more natural vocabulary breadth.

4. Syntactic & Lexical Complexity:

Different LLMs employ distinct complexity strategies: Llama shows the highest syntactic complexity (parse depth 7.41) using deeply nested sentence structures, while Claude and GPT-4 favor lexical complexity with longer, more formal words (5.05 and 4.97 chars) in moderately complex structures (parse depth ~5.3). Human text with the lowest values on both dimensions (parse depth 4.72, word length 4.45), reflecting a natural, conversational style prioritizing simplicity and directness. Interestingly, Qwen mirrors human patterns more closely, suggesting it may be optimized for accessible, casual communication rather than sophisticated formality.

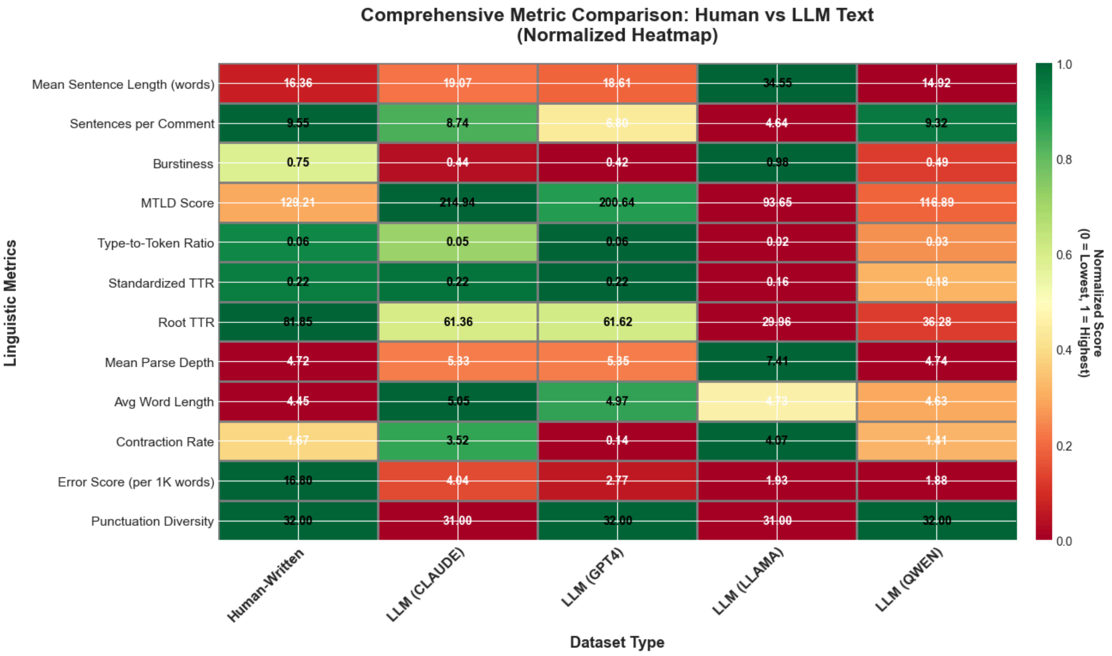
5. Formality and Tone (Contractions):

GPT-4 shows a near-zero contraction rate (0.14%), indicating an extremely formal, distinct "voice." In contrast, Humans (1.67%) and Claude (3.52%) use contractions more freely, creating a more conversational tone.

Summary:

Human text on Reddit is characterized by higher error rates, moderate vocabulary complexity, and natural structural variation. Advanced LLMs (Claude, GPT-4) are distinguishable by their "cleanliness," high lexical density, and overly consistent sentence structures.

5. Appendix - Metrics Heat Map ***More Visualizations available in the notebook**


Comprehensive Metric Comparison: Human vs LLM Text
(Normalized Heatmap)
Linguistic Metrics
Mean Sentence Length (words)
Sentences per Comment
Burstiness
MTLD Score
Type-to-Token Ratio
Standardized TTR
Root TTR
Mean Parse Depth
Avg Word Length
Contraction Rate
Error Score (per 1K words)
Punctuation Diversity
Human-Written
LLM (CLAUDE)
LLM (GPT4)
LLM (LLAMA)
LLM (QWEN)
Dataset Type
Normalized Score
(0 = Lowest, 1 = Highest)