HLU: Human Vs LLM Generated Text Detection Dataset for Urdu at Multiple Granularities

Iqra Ali, Jesse Atuhurra, Hidetaka Kamigaito, Taro Watanabe
{ali.iqra.ai6, atuhurra.jesse.ag2, kamigaito.h, taro}@is.naist.jp
Nara Institute of Science and Technology

Abstract

The rise of large language models (LLMs) generating human-like text has raised concerns about misuse, especially in low-resource languages like Urdu. To address this gap, we introduce the HLU dataset, which consists of three datasets: Document, Paragraph, and Sentence-level. The document-level dataset contains 1,014 instances of human-written and LLM-generated articles across 13 domains, while the paragraph and sentence-level datasets each contain 667 instances. We conducted both human and automatic evaluations. In the human evaluation, the average accuracy at the document level was 36.5%, while at the paragraph and sentence levels, accuracies were 76.95% and 82.09%, respectively. For automatic evaluation, we fine-tuned the XLM-RoBERTa model for both monolingual and multilingual settings achieving consistent results in both. Additionally, we assessed the performance of GPT-4 and Claude-3-Opus using zero-shot prompting. Our experiments and evaluations indicate that distinguishing between human and machine-generated text is challenging for both humans and LLMs, marking a significant step in addressing this issue in Urdu. The evaluation, dataset [1], and code [2] is accessible publicly for research purpose.

1 Introduction

The increased usage of large language models (LLMs) is aiding the generation of text that closely resembles human writing (Zheng et al., 2023; Abdullin et al., 2024). Whereas LLMs demonstrate remarkable proficiency in various natural language processing (NLP) tasks (Nozza et al., 2022); (Xiao and Chen, 2023); (Potter and Palmer, 2023) and (Mitchell et al., 2023), their potential misuse for spreading misinformation (Chen and Shu, 2023),

Human-Written Text	GPT-3.5 Generated Text	GPT-4 Generated Text
کرکٹ ایک ایسا کھیل ہے جو گیارہ کھلاڑیوں پر مشتمل دو ٹیموں کے درمیان میں کھیلا جاتا ہے۔ یہ کھیل گیند اور بلے کے ذریعے کھیلا جاتا ہے جس کا میدان بیضوی شکل کا ہوتا ہے۔ میدان کے درمیان میں 20.12 میٹر (22 گز) کا مستقر بنا ہوتا ہے جسے پچ کہا جاتا ہے۔	کرکٹ پاکستان کی جموں اور کشمیر کے علاقوں میں پسند کیا جانے والا ایک مقبول کھیل ہے۔ یہ کھیل نے پاکستان کی قومی ورزش کی حیثیت حاصل کی ہے اور لوگوں کے دلوں میں اہم جگہ رکھتا ہے۔ کرکٹ نے پاکستان کی جموں اور کشمیر کے علاقوں میں تیزی سے فیم پایا ۔	کرکٹ، جو 16ویں صدی کے آخر میں جنوبی انگلینڈ میں اپنی جڑیں مضبوط کرنے والا ایک کھیل ہے، آج عالمی سطح پر ایک جنون بن چکا ہے۔ اس کھیل کی تاریخ، جذبہ، اور انفرادیت نے اسے صرف ایک کھیل سے بڑھ کر ایک ثقافتی عنصر بنا دیا ہے، جس کے پرستار دنیا بھر میں پائے جاتے ہی۔

Figure 1: Overview of our proposed dataset.

and disrupting educational systems (Wang et al., 2023) has raised immediate concerns. Even, In the banking sector, traditional machine learning based models, such as those described in (Aurna et al., 2024), can be effectively replaced by the adoption of large language models (LLMs). In paraphrase generation tasks, machines can detect paraphrased text (Ali et al., 2024), yet identifying the source of such text remains uncertain. This underscores the urgent need to develop automated systems capable of detecting machine-generated text (Mitchell et al., 2023; Wang et al., 2023). Recent studies have shown that even humans struggle to differentiate between human-generated text (HGT) and machine-generated text (MGT) (Gehrmann et al., 2019); (Singh et al., 2024).

In the past the datasets and models have been developed for high-resource languages like English (Hasan et al., 2021), Japanese (Wang et al., 2024a; Das et al., 2024), Chinese (Guo et al., 2023a) and more. Yet, the low-resource languages such as Urdu remain under-explored. Moreover, previous work has primarily focused on document-level detection, with little focus on fine-grained detection at the sentence and paragraph levels, which is crucial given users' tendency to modify parts of texts using LLMs (Mitchell et al., 2023; Wang et al., 2023).

To address the above gaps, we introduce the HLU dataset. (See Figure 1)

- We proposed three datasets across 13 domains:

[1] https://huggingface.co/datasets/iqraali/Urdu-HumanvsMachine-Dataset
[2] https://huggingface.co/iqraali/humanvsllm

Proceedings of the 31st International Conference on Computational Linguistics, pages 3495–3510
January 19–24, 2025.

Document-level (1,014 instances), Paragraph-level (667), and Sentence-level (667) for human-written and LLM-generated text.

- Human evaluation revealed average accuracy of 36.5% at the document level, 76.95% at the paragraph level, and 82.09% at the sentence level, highlighting the difficulty for humans in distinguishing between human and machine-generated text.
- We finetuned the XLM-RoBERTa achieving F1 scores of 84.5% (paragraph) and 70.6% (sentence).We can see that when given more data the model performs better but human performs best when given short texts.

2 Related Work

Ongoing efforts aim to develop datasets for detecting human vs LLM-generated text (see Table 1).

Human Detection: The Turing Test (Oppy and Dowe, 2003) is used to evaluate chatbot responses by assessing whether texts are human or machine-generated, a standard method for evaluating generation systems (Lowe et al., 2017; Van Der Lee et al., 2019; Kreps et al., 2022; Gehrmann et al., 2019). Van Der Walt and Eloff (2018) emphasize the need for human ability to identify fake content.

Automatic Detection: DetectGPT (Mitchell et al., 2023) assumes LLM-generated texts have lower model probability than human texts, while Turnitin (Batane, 2010) and black-box detection utilize API-level access for classification (Dugan et al., 2020). Guo et al. (2023a) explored text characteristics through fine-tuning pre-trained models on question-and-answer datasets.

3 Dataset Creation

The data creation process is outlined in subsections.

3.1 Data Source

We used Wikipedia[4] (see Figure 3 in Appendix A), for our data collection. To construct our LLM-generated part, we utilized GPT-3.5-turbo and GPT-4-turbo via API. LLM access via API was preferred due to concerns regarding potential data leakage and privacy issues (Balloccu et al., 2024).

3.2 *Document-level* Data

Human-written Data: After deciding the source of the data, we diversified our dataset with 13 categories because articles about these categories are widely available in Wikipedia. We manually collected 338 human-written articles ***26 per category***, written by humans to build our corpus over a duration of 2 to 3 months. (see Table 8 in Appendix B)

Prompting for LLM-generated Data: After finalizing the categories, we leveraged in-context learning (Brown et al., 2020) and prompt engineering with GPT-3.5-turbo and GPT-4-turbo to create an LLM-generated corpus. This process involved iteratively refining the prompts through a cycle of design, execution, and analysis of outputs until we derived prompts that effectively met the requirements of our study. We collected 338 articles each by prompting GPT-3.5-turbo and GPT-4-turbo for 13 categories. The prompt is presented in Table **??**.

Self-Criticism: It has gained attention in recent years (Tan et al., 2023; Peng et al., 2023; Asai et al., 2023). We prompted GPT-3.5-turbo and GPT-4-turbo to engage in self-critique of their generated textual outputs. We evaluated four key areas of potential error (Dugan et al., 2020) (see **Q1-Q6** in Table 2). Details in Appendix D.

3.3 *Paragraph, Sentence Level* Data

For building the *paragraph-level* dataset, we utilized our document-level dataset consisting of 1,014 entries. We methodically extracted groups of sentences, each comprising three or more sentences, from our articles to create paragraphs, resulting in 667 instances for our paragraph-level dataset. To create the dataset at the *sentence level*, we treat one or more words expressing a complete thought, statement, question, or command as individual sentences, resulting in a total of 667 instances.

3.4 Data Statistics

Table 9 in Appendix C depicts the word counts from 22,051 (sentence level) to 110,965 (document level). Vocabulary richness (TTR) improves from 0.22 to 0.71 and maximum word counts range from 91 (sentence) to 738 (document).

3.5 Example of Text from Human, GPT-3.5 and GPT-4

The example text in Table 3 shows the unique differences produced by humans, GPT-3.5, and GPT-4.

4 Evaluation

We have performed two types of evaluation: human evaluation and automatic evaluation.

[4] https://en.wikipedia.org/wiki/Urdu

Corpus Name	Size	Source	Language	Domain
HC3 (Guo et al., 2023a)	84K (train)	ChatGPT	En	Q/A, CS.
M4 (Wang et al., 2024b)	147K	Multiple	Multiple	General
GPT-2 Output [3]	250K (train)	GPT-2	En	Web
GPABenchmark (Liu et al., 2023)	1,200K (train)	GPT-3.5	En	SW
Deepfake (Altuncu et al., 2022)	319K (train)	GPT, LLaMA, etc.	En	News, Q/A, etc.
HLU (Ours)	**1,014**	**Wikipedia, GPT-3.5, GPT-4**	**Ur**	**General (13 categories)**

Table 1: Existing Human vs LLM-generated text detection dataset.

Input Prompt in English and Urdu
Please generate a short story in the Urdu language, describing the life of a fictional character. Use simple yet engaging language and avoid overly complex vocabulary.
سادہ لیکن دلکش زبان استعمال کریں اور زیادہ پیچیدہ الفاظ سے گریز کریں۔ براہ کرم اردو زبان میں ایک مختصر کہانی تیار کریں، جو ایک خیالی کردار کی زندگی کو بیان کرے۔
Q1. How effectively does the generated text convey the emotions and personality of the character?
1- تیار کردہ متن کردار کے جذبات اور شخصیت کو کتنا مؤثر طریقے سے پیش کرتا ہے؟
Q2. On a scale of 1 to 5, how well does the story maintain coherence from beginning to end?
2- 1 سے 5 کے پیمانے پر، کہانی کا آغاز سے آخر تک تسلسل کس حد تک برقرار رہتا ہے؟
Q3. Does the generated text reflect cultural and contextual understanding?
3- کیا تیار کردہ متن ثقافتی اور سیاق و سباق کی تفہیم کو ظاہر کرتا ہے؟
Q4. Are there any repetitive phrases or ideas in the text?
4- کیا متن میں کوئی دہرائے گئے جملے یا خیالات ہیں؟
Q5. Are there any logical inconsistencies in the events or character actions?
5- کیا واقعات یا کردار کے اعمال میں کوئی منطقی تضادات ہیں؟
Q6. Does the text meet the requirements of being simple and engaging as instructed?
6- کیا متن سادہ اور دلکش ہونے کی ہدایات پر پورا اترتا ہے؟

Table 2: Prompt used to evaluate LLM-generated Urdu stories

Type	Text in Urdu with English Translation
Human	استاد نصرت فتح علی خان عظیم پاکستانی قوال، موسیقار، موسیقی ڈائریکٹر اور بنیادی طور پر قوالی کے گلوکار تھے۔ Ustad Nusrat Fateh Ali Khan was a great Pakistani Qawwal, composer, music director, and mainly Qawwali singer.
GPT-3.5	نصرت فتح علی خان پاکستان کی معروف گلوکاری کے شہنشاہ تھے جنہوں نے اپنی زندگی بھر موسیقی کی دنیا میں خود کو قائم کیا۔ Nusrat Fateh Ali Khan was the famous singing emperor of Pakistan who established himself in the world of music throughout his life.
GPT-4	نصرت فتح علی خان، جنہیں قوالی کے شہنشاہ کے طور پر جانا جاتا ہے، پاکستان کے معروف گلوکار تھے جنہوں نے قوالی موسیقی کو عالمی سطح پر مقبول بنایا۔ Nusrat Fateh Ali Khan, known as the Emperor of Qawwali, was a renowned Pakistani singer who popularized Qawwali music globally.
Analysis	The example text from our corpus demonstrates the degrees of emotional expression, engagement, and language complexity between humans, GPT-3.5, and GPT-4. The human text expresses genuine emotion and engagement, the GPT-3.5 generated text lacks depth and emotional touch. In contrast, the GPT-4 text showcases a higher level of sophistication and language complexity, reflecting its ability to produce detailed content.

Table 3: Comparison of text generation between humans and GPT models in Urdu

4.1 Human Evaluation

The human evaluation is based on the **Turing Test** at *document* level. The Turing Test (Oppy and Dowe, 2003) tests a machine's ability to exhibit intelligent behavior that is indistinguishable from a human. We invited two evaluators: one who has used, and one who has not used GPT. To conduct the human evaluation, we introduced three sub-tasks.

Task-1: In this task, human annotators are given *one article* at a time and they have to guess if it has been written by humans, GPT3.5 or GPT4

Task-2: We frame the detection problem as a binary classification task over *one article*: given an excerpt from a text, label it as either human-written or LLM-generated.

Human vs GPT-3.5: Human annotators are provided with two articles from the same category at a time, and they have to guess which article is from human and which one is from GPT-3.5.

Human vs GPT-4: Similarly, human annotators are provided with two articles from the same category at a time, and they have to guess which one is from human and which one is from GPT-4.

Task-3: Human annotators are provided with *three articles* at a time, and they have to determine which one is from human and which one is from GPT-3.5 and GPT-4.

4.1.1 Annotation Guidelines

Two native Urdu-speaking annotators, aged 20-30 with backgrounds in computer science and bioscience were tasked with annotating the documents, paragraphs, and sentences. The annotation guidelines were adapted from previous works (Dugan et al., 2020; Wu et al., 2024) for consistency and accuracy in the evaluation process. These guidelines were provided to the annotators to ensure a clear understanding of how to evaluate the different types of text.

- Human-written text: The annotators were advised to evaluate human-written text, which is ide-

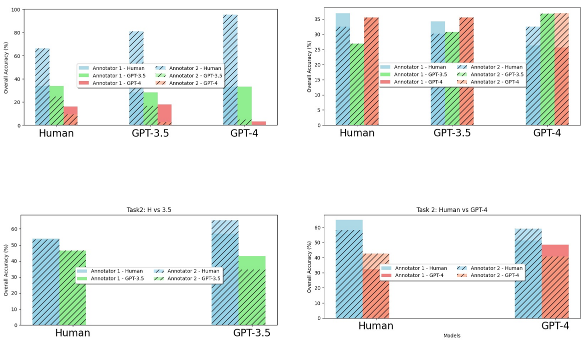

Figure 2: Comparison of accuracy provided by humans for Task 1, Task 3 and Task 2. All evaluations are at *document* level.

ally characterized by the nuances of human thought processes and creativity. Such text mostly includes idiosyncrasies, personal narratives, story and an individual style that reflects the author's experiences and perspectives. Human authors tend to exhibit inconsistencies, shifts in tone, and occasional errors, which convey a natural, less structured approach to writing. Additionally, human-written text demonstrates a deep understanding of context, cultural references, and emotional intelligence, making it contextually rich and adaptable.

- Machine-generated text: In contrast, machine-generated text, created through sophisticated algorithms like GPT-3.5 and GPT-4, can imitate human writing styles and show high levels of coherence and grammatical correctness. However, the annotators were briefed to remember that such text may lack creativity and a personal touch. Machine-generated text tends to be very consistent, maintaining a uniform style and tone, but it may also present inconceivable facts or awkward phrasing due to its reliance on statistical patterns rather than real-world understanding. Furthermore, machine-generated text can struggle with tasks that require deep contextual, cultural or emotional knowledge.

4.1.2 Document-level Evaluation Results

IAA Results: The IAA score between two evaluators per task, measured by Cohen's Kappa (Cohen, 1960), is shown in Table 4.

Task1	Task2: HvsGPT-3.5	Task2: HvsGPT-4	Task3
0.0273	0.0563	-0.0098	0.0242

Table 4: Document-level IAA scores.

Human Evaluation Accuracy – Document level: Figure 2 demonstrates that *annotator 1*, who was familiar with ChatGPT, achieved higher accuracy than *annotator 2*. Human evaluators found it challenging to distinguish between GPT-generated and human-written texts, with an average document-level accuracy of 36.5%, only slightly better than random guessing. Details in Appendix F.

4.1.3 Paragraph, Sentence Level Evaluation Results

IAA Results: Annotator agreement at both the paragraph and sentence-level is shown in Table 5.

IAA	GPT-3.5	GPT-4
Paragraph	0.7568	0.6993
Sentence	0.8845	0.7259

Table 5: Paragraph and sentence-level IAA scores.

Human Evaluation Accuracy – Paragraph and Sentence level: We performed human evaluation for Urdu achieving accuracy of *81.7%* and *82.6%* at sentence and paragraph level.

Task	Size	F1 Scores	
		Paragraph	Sentence
HvsGPT-3.5	667	0.7695	0.7272
HvsGPT-4	667	0.8179	0.8209

Table 6: Performance at Paragraph and Sentence levels.

4.1.4 Discussion

The document, Paragraph and Sentence level results indicating that as the text length increases, it becomes more difficult for humans to differentiate between human and machine-generated content. This also suggests that GPT's text generation for Urdu is highly advanced, as human annotators struggled to distinguish between texts.

4.2 Automatic Evaluation

We deployed XLM-RoBERTa[5] (Conneau et al., 2020), in our experiments. The batch size is set to 16, the number of epochs to 5, and the learning rate to $1e-5$.

4.2.1 Monolingual Settings

Binary classification (Our data): At *paragraph-level* and *sentence-level*, we trained a XLM-RoBERTa classifier to distinguish between *Human*

[5]https://huggingface.co/xlm-roberta-base

Classification	Data (%)	Size	F1-score
HLU- Ours – Only Ur			
Paragraph			
Human vs GPT-3.5	100	667	**0.8451**
	75	508	0.8246
	50	338	0.6410
Human vs GPT-4	100	667	**0.8451**
	75	508	0.8421
	50	338	0.7842
Sentence			
Human vs GPT-3.5	100	667	**0.7035**
	75	508	0.6842
	50	338	0.6619
Human vs GPT-4	100	667	**0.7655**
	75	508	0.7193
	50	338	0.6763
HC3 (Guo et al., 2023b) – Only En			
Human vs M	–	667	0.7478
	–	24,321*	0.9500
MULTITuDE (Macko et al., 2023) – w/o Ur			
Human vs GPT-3.5	–	667	0.6302
Human vs GPT-3.5	–	16,292*	0.8444
Human vs GPT-4	–	667	**0.7728**
Human vs GPT-4	–	16,292*	0.8687
M4 (Wang et al., 2024b) – w/ Ur			
Human vs M	–	667	**0.7520**
Human vs M	–	16,000*	0.8652

Table 7: Automatic evaluation at monolingual and multilingual settings by fine-tuning *XLM-RoBERTa*.

and GPT-3.5, and *Human and GPT-4* texts, respectively. We used the training data of our Urdu corpus at data regimes: 50%, 75%, and 100%.

Binary Classification (En): For the English language comparison, we fine-tuned XLM-RoBERTa on the dataset introduced by Guo et al. (2023b), referred to as *HC3*. Initially, we fine-tuned XLM-RoBERTa on 667 instances, followed by an the whole dataset 24,321. This process resulted in F1 scores of 74% and 95%, respectively.

4.2.2 Multilingual Settings

Binary Classification (MULTITuDE): For the multilingual comparison, we fine-tuned XLM-RoBERTa using *MULTITuDE* Macko et al. (2023) dataset which includes: en, es, ru, nl, ca, cs, de, zh, pt, ar, uk. To ensure a fair comparison, we specifically used data generated by GPT-3.5-turbo and GPT-4, along with human-written instances. In our experiments we used two data regimes *667* and *16,292*. The F1 scores are 63.02% for *Human vs GPT-3.5* which shows a decline in F1 score as compared to our dataset, at *sentence-level*.

Binary Classification (M4): In our experiments we compared our dataset, HLU, with the multilingual M4 (Wang et al., 2024b) dataset which includes Urdu. The Urdu text in M4 was created using *gpt-3.5-turbo-instruct* version *davinci-text-003*. The F1 scores are 75.20% and 86.52% which shows given more data the model F1 scores improves. Our experiment results with M4 not only emphasize the observation that more data leads to better text classification but the results are consistent in both the monolingual and multilingual settings. This finding emphasizes the contribution of our proposed dataset.

4.2.3 Zero-shot Prompting of LLM:

We performed evaluation with LLM, i.e., GPT-4o and Claude-3-Opus, via zero-shot prompting but both LLM struggled to distinguish between human-written and LLM-generated texts. See Appendix G for prompt.

4.2.4 Results:

Table 7 shows that XLM-RoBERTa classifier performs best when more data is given. In all cases, XLM-RoBERTa performed best for 100% data, and the performance decreases with decreasing data size. We can see that when given more data the model performs better but human performs best when given short sentences with the F1 score of 82% at *sentence level* in Urdu. Yet, XLM-RoBERTa fine-tuned on English data performs better with more context, but given the same data (667) as ours, it performed worse than Urdu.

5 Conclusion

Our findings indicate that it is difficult for humans to differentiate between human vs LLM-generated texts. The average accuracy under human evaluation at document level is *36.5%*, which shows that it is difficult for humans to differentiate between human vs LLM-generated text as the text-generation ability of GPT is good for Urdu language.

For the automatic evaluation, when XLM-RoBERTa is fine-tuned, F1-scores are *84.5%* and *70.6%* at paragraph and sentence level for Urdu. However, F1-scores for human evaluation at paragraph and sentence level are *81.7%* and *82.6%*. Thus, it is increasingly important to develop corpora for underrepresented languages to train the text-detection classifiers and our evaluations highlights that we need more of such datasets to train models for text detection tasks in Urdu.

6 Limitations

6.1 Limited Annotators

Our human evaluation was conducted with only two annotators, which may not fully represent the diversity of human judgment in detecting human vs LLM-generated text.

6.2 Models from the Same Family

All models used for text generation are from the same family, i.e. GPT. The lack of support for Urdu in most of the LLMs holds a major challenge for performing automatic evaluations in Urdu.

6.3 Controlled Length

For our task we used length constraints during text generation using GPT-3.5 and GPT-4. Controlled-length summarization (Juseon-Do et al., 2024) is important for producing concise and precise text adjusted to specific requirements, such as space-limited contexts or user-defined preferences. However, models often struggle with length control due to limited model capabilities and the absence of inherent mechanisms for constraint handling.

6.4 Adoption of Single Template

We used a single-template approach for generating text. As discussed in (Sakai et al., 2024) using multiple templates ensures a more robust and fair evaluation of model capabilities by considering the variability in performance across different prompts, our decision to use a single template is driven by specific constraints like uniformity in data collection.

References

Yelaman Abdullin, Diego Molla-Aliod, Bahadorreza Ofoghi, John Yearwood, and Qingyang Li. 2024. Synthetic dialogue dataset generation using llm agents. *Preprint*, arXiv:2401.17461.

Iqra Ali, Hidetaka Kamigaito, and Taro Watanabe. 2024. Monolingual paraphrase detection corpus for low resource Pashto language at sentence level. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11574–11581, Torino, Italia. ELRA and ICCL.

Enes Altuncu, Virginia N. L. Franqueira, and Shujun Li. 2022. Deepfake: Definitions, performance metrics and standards, datasets and benchmarks, and a meta-review. *Preprint*, arXiv:2208.10913.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Nahid Ferdous Aurna, Md Delwar Hossain, Latifur Khan, Yuzo Taenaka, and Youki Kadobayashi. 2024. Fedfusion: Adaptive model fusion for addressing feature discrepancies in federated credit card fraud detection. *IEEE Access*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Tshepo Batane. 2010. Turning to turnitin to fight plagiarism among university students. *Journal of Educational Technology & Society*, 13(2):1–12.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Canyu Chen and Kai Shu. 2023. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Debarati Das, Karin De Langis, Anna Martin, Jaehyung Kim, Minhwa Lee, Zae Myung Kim, Shirley Hayati, Risako Owan, Bin Hu, Ritik Parkar, et al. 2024.

Under the surface: Tracking the artifactuality of llm-generated data. *arXiv preprint arXiv:2401.14698*.

Liam Dugan, Daphne Ippolito, Arun Kirubarajan, and Chris Callison-Burch. 2020. Roft: A tool for evaluating human detection of machine-generated text. *arXiv preprint arXiv:2010.03070*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023a. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023b. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arxiv:2301.07597*.

Tahmid Hasan, Abhik Bhattacharjee, Md Saiful Islam, Kazi Samin, Yuan-Fang Li, Yong-Bin Kang, M Sohel Rahman, and Rifat Shahriyar. 2021. Xl-sum: Large-scale multilingual abstractive summarization for 44 languages. *arXiv preprint arXiv:2106.13822*.

Juseon-Do Juseon-Do, Hidetaka Kamigaito, Manabu Okumura, and Jingun Kwon. 2024. InstructCMP: Length control in sentence compression through instruction-based large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8980–8996, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Zeyan Liu, Zijun Yao, Fengjun Li, and Bo Luo. 2023. Check me if you can: Detecting chatgpt-generated academic writing using checkgpt. *Preprint*, arXiv:2306.05524.

Ryan Lowe, Michael Noseworthy, Iulian V Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic turing test: Learning to evaluate dialogue responses. *arXiv preprint arXiv:1708.07149*.

Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. Multitude: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Debora Nozza, Federico Bianchi, Anne Lauscher, Dirk Hovy, et al. 2022. Measuring harmful sentence completion in language models for lgbtqia+ individuals. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Graham Oppy and David Dowe. 2003. The turing test.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Lucas Potter and Xavier-Lewis Palmer. 2023. Post-llm academic writing considerations. In *Proceedings of the Future Technologies Conference*, pages 154–163. Springer.

Yusuke Sakai, Adam Nohejl, Jiangnan Hang, Hidetaka Kamigaito, and Taro Watanabe. 2024. Toward the evaluation of large language models considering score variance across instruction templates. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 499–529, Miami, Florida, US. Association for Computational Linguistics.

Aakash Singh, Deepawali Sharma, Abhirup Nandy, and Vivek Kumar Singh. 2024. Towards a large sized curated and annotated corpus for discriminating between human written and ai generated texts: A case study of text sourced from wikipedia and chatgpt. *Natural Language Processing Journal*, 6:100050.

Xiaoyu Tan, Shaojie Shi, Xihe Qiu, Chao Qu, Zhenting Qi, Yinghui Xu, and Yuan Qi. 2023. Self-criticism: Aligning large language models with their understanding of helpfulness, honesty, and harmlessness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 650–662.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 355–368.

Estée Van Der Walt and Jan Eloff. 2018. Using machine learning to detect fake identities: Bots vs humans. *IEEE Access*, 6:6540–6549.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao

Dong, Ming Ding, and Jie Tang. 2024a. Cogvlm: Visual expert for pretrained language models. *Preprint*, arXiv:2311.03079.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Alham Fikri Aji, and Preslav Nakov. 2023. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. *Preprint*, arXiv:2305.14902.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4: Multi-generator, multi-domain, and multilingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A survey on llm-generated text detection: Necessity, methods, and future directions. *Preprint*, arXiv:2310.14724.

Le Xiao and Xiaolin Chen. 2023. Enhancing llm with evolutionary fine tuning for news summary generation. *arXiv preprint arXiv:2307.02839*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.

A Appendix

We collected the human written articles from Wikipedia and example is shown in Figure 3.

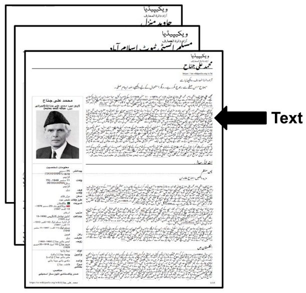

Figure 3: An exemplar article in Urdu Wikipedia.

B Human Written Data Categories

The dataset contains 13 categories as mentioned in Table 8.

Category	**# of Articles**
1- Famous Personality	26 ✗ 3
2- Sports	26 ✗ 3
3- Health	26 ✗ 3
4- Science and Technology	26 ✗ 3
5- Events	26 ✗ 3
6- Crime and Law	26 ✗ 3
7- History	26 ✗ 3
8- Countries	26 ✗ 3
9- Hobbies	26 ✗ 3
10- Food	26 ✗ 3
11- Movies	26 ✗ 3
12- Natural Disaster	26 ✗ 3
13- Animals	26 ✗ 3
	Total = 338 + 338 + 338 = 1,014

Table 8: Categories and the number of articles collected from Wikipedia, GPT-3.5, and GPT-4.

C Combined Dataset Characteristics at Sentence, Paragraph, and Document Levels

The Table 9 shows that our corpus is diverse, making it a realistic, robust resource for the human vs LLM generated text-detection task.

D Criteria of Self-Criticism

To explore the phenomena of self-criticism in large language models (LLMs) like GPT-3.5 and GPT-4, we prompted these models to self-critique their generated articles. The process was evaluated on four key areas of potential error or improvement: repetition, common-sense errors, factual errors, and incoherence and topic drift.

Metric	Level	Count		
Total words	Sentence Paragraph Document	22,051 35,051 110,965		
Total unique words	Sentence Paragraph Document	4,895 11,399 17,567		
Average length	Sentence Paragraph Document	23.97 30.15 168.15		
Vocabulary richness (TTR)	Sentence Paragraph Document	0.22 0.51 0.71		
Label	**Level**	**Human**	**GPT-3.5**	**GPT-4**
Max no. of words	Sentence Paragraph Document	91 120 467	85 85 677	98 98 738
Min no. of words	Sentence Paragraph Document	4 21 35	7 15 49	6 18 58
Mean of words	Sentence Paragraph Document	22.4 22.7 22.7	21.2 21.9 21.9	21.4 21.9 21.9
Median of words	Sentence Paragraph Document	21 21 41	20 20 37	20 20 38
Vocabulary Size (V)	Sentence Paragraph Document	1,007 2,857 2,857	1,174 2,174 2,174	1,088 2,088 2,088
Density (D)	Sentence Paragraph Document	23.6 37.6 37.6	19.2 30.2 30.2	21.7 28.7 28.7

Table 9: Data Characteristics at Sentence, Paragraph, and Document Levels

Question	GPT-4 Self-Criticism Response
Q1	انسانی تحریر کی مماثلت: یہ جواب عبدالستار ایدھی کی زندگی اور ان کے کام کی تفصیلات پیش کرتا ہے، جو انسانی تحریر کی طرح سلیس اور معلوماتی ہے۔ تاھم، اس میں مزید انفرادی تفصیلات اور ذاتی قصے شامل کیے جا سکتے ھیں تا کہ انسانی ربط اور گہرائی میں اضافہ ھو۔ Similarity of Human Writing: This answer details the life and work of Abdul Sattar Edhi, which is as smooth and informative as human writing. However, it can include more individual details and personal anecdotes to add human connection and depth. بہتری کی گنجائش: متن میں ایدھی صاحب کے کچھ خاص اقدامات یا ان کے ذاتی تجربات کی مزید مثالیں شامل کی جا سکتی ھیں تا کہ پڑھنے والے کو ان کی شخصیت کا زیادہ قریب سے احساس ھو۔ Room for improvement: The text could have included more examples of some of Edhi Sahib's specific actions or his personal experiences so that the reader gets a closer sense of his personality.
Q2	اعتماد کی سطح (1 سے 5 پیمانے پر): میرا اعتماد اس جواب پر 4 ہے۔ اس میں معلومات کی درستگی اور تفصیل کا خیال رکھا گیا ہے، لیکن ھمیشہ بہتری کی گنجائش رھتی ہے۔ Confidence level (on a scale of 1 to 5): My confidence is 4 on this answer. Care has been taken to ensure the accuracy and detail of the information, but there is always room for improvement.
Q3	متن میں کوئی نمایاں تکرار نہیں ہے۔ ھر پیراگراف مختلف پہلوؤں کو اجاگر کرتا ہے۔ There are no significant repetitions in the text. Each paragraph highlights a different aspect.
Q4	متن میں کوئی واضح عام فہم غلطیاں نہیں پائی جاتیں۔ There are no obvious common sense errors in the text.
Q5	جہاں تک میرے علم کی حد تک ہے، متن میں کوئی حقیقتی غلطیاں موجود نہیں ھیں۔ To the best of my knowledge, there are no factual errors in the text.
Q6	متن میں کوئی بے ربطی یا موضوع کی خرابی نہیں پائی جاتی۔ یہ عبدالستار ایدھی کی زندگی اور ان کے کام کو مؤثر طریقے سے پیش کرتا ہے۔ There is no incoherence or distortion of subject matter in the text. It effectively presents the life and work of Abdul Sattar Edhi.

Table 10: GPT-4 Self-Criticism Response

Repetition: We prompted the models to assess its response for any redundant use of words or phrases, or unnecessary copy of information.

Common-sense errors: The model was prompted to review its answers for errors that violate basic rules of logic, general knowledge, or common understanding.

Factual errors: The model was prompted to verify its accuracy of the information it provided, identifying any inaccuracies or falsehoods.

Incoherence and topic drift: The model was asked to evaluate the logical flow of its answers, as well as to check for sudden shift in subject matter without clear transitions.

D.1 Example of GPT-3.5 and GPT-4 Self-Criticism Responses

See Table 10 and 11 for example of self- criticism from GPT-3.5 and GPT-4.

D.2 Self-Criticism Results

In comparing the self-criticism responses between GPT-4 and GPT-3.5, we found that GPT-4 is good in providing detailed critiques of its own answers as compared to GPT-3.5.

GPT-4 response displays specific details and considerations, showcasing a deep level of analysis and mirroring on the clarity, completeness, and purpose of its generated content. While, GPT-3.5's response is more general and less specific, indicating a lower level of self-criticism compared to GPT-4.

E Implementation Settings

Model Size Table 12 shows the model size parameters.

Question	GPT-3.5 Self-Criticism Response
Q1	یہ مواد انسانی تحریر سے بہت مماثل ہے کیونکہ اس میں موجودہ وضاحت کی شکل میں جانوروں کی معلومات فراہم کی گئی ہیں۔ This material is very similar to human writing in that it provides animal information in the form of an existing description.
Q2	میں اپنے جواب کے بارے میں 4 پر اعتماد رکھتا ہوں۔ I am confident about my answer 4.
Q3	نہیں، متن میں کوئی تکرار نہیں ہے۔ No, there is no repetition in the text.
Q4	متن میں عام فہم غلطیاں نہیں ہیں۔ There are no common sense errors in the text.
Q5	متن میں کوئی حقیقتی غلطی نہیں ہے۔ There are no factual errors in the text.
Q6	متن میں بے ربطی نہیں ہے، لیکن اگر کچھ مزید تفصیلات شامل کی جائیں تو موضوع کو مزید واضح بنایا جا سکتا ہے۔ The text is not incoherent, but the subject could be made clearer if a few more details were added.

Table 11: GPT-3.5 Self-Criticism Response

Model	XLM-Roberta
Parameters	279M

Table 12: Specifications of the XLM-Roberta Model

Hyperparameters: We have used the hyperparameters defined in Table 13.

Parameter	Value
Batch Size	16
Optimizer	Adam
# Epochs	5
Sequence Length	50
Learning Rate	1e-5

Table 13: Training Parameters

Parameters For Packages We have used the packages showcased in Table 14.

F Human Evaluation

From the Figure 4, 5, we can see the overall heat-map of Annotator 1 and 2 across Task 1, 2 and 3. Whereas Figure 6, 7, 8, 9, shows the heat-map across all 13 categories of Annotator 1 and 2. The Figure 10, 11 is the bar-plot of overall number of correct Labels, number of wrong Labels, 1 correct 2 wrong Labels for Annotator 1 and 2. Figure 12 shows the comparison between Annotator 1 and Annotator 2.

Packages	Version
Transformers	4.41.2
Scikit-learn	1.2.2
Seaborn	0.12.2
Matplotlib	3.7.5

Table 14: Parameters For Packages

G Zero-Shot Prompting for GPT-4o and Claude-3-Opus

We performed evaluation with LLM, i.e., GPT-4o and Claude-3-Opus via API i.e.,gpt-4-turbo-2024-04-09 and claude-3-opus-20240229 respectively, via zero-shot prompting but both LLM struggled to distinguish between human-written and LLM-generated texts. See Table 15.

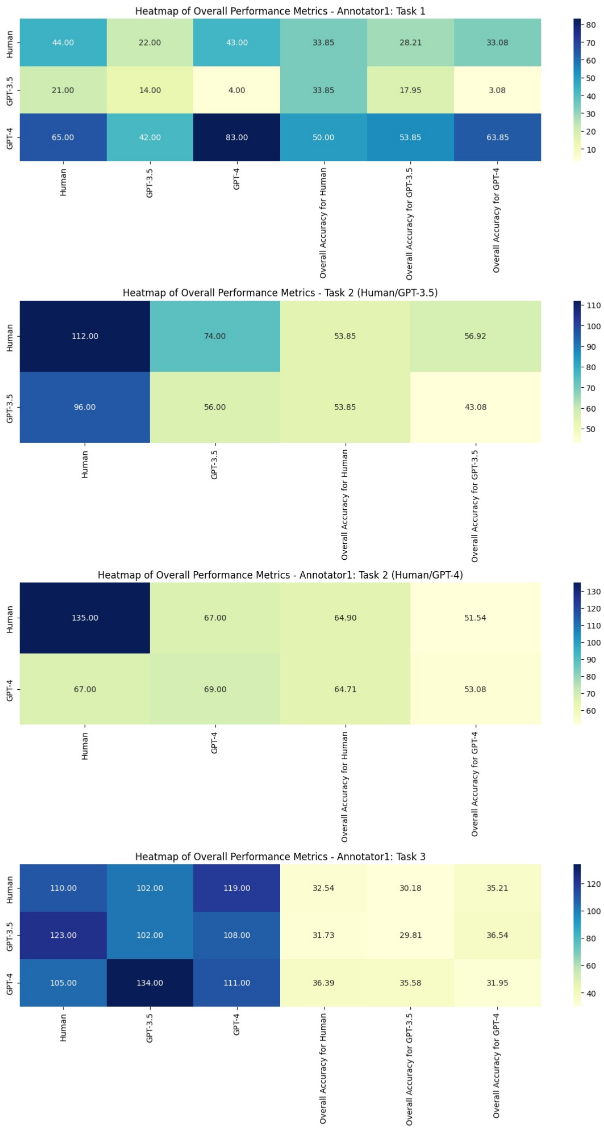

Figure 4: Heatmap of Annotator 1 across Task1, Task2, Task3

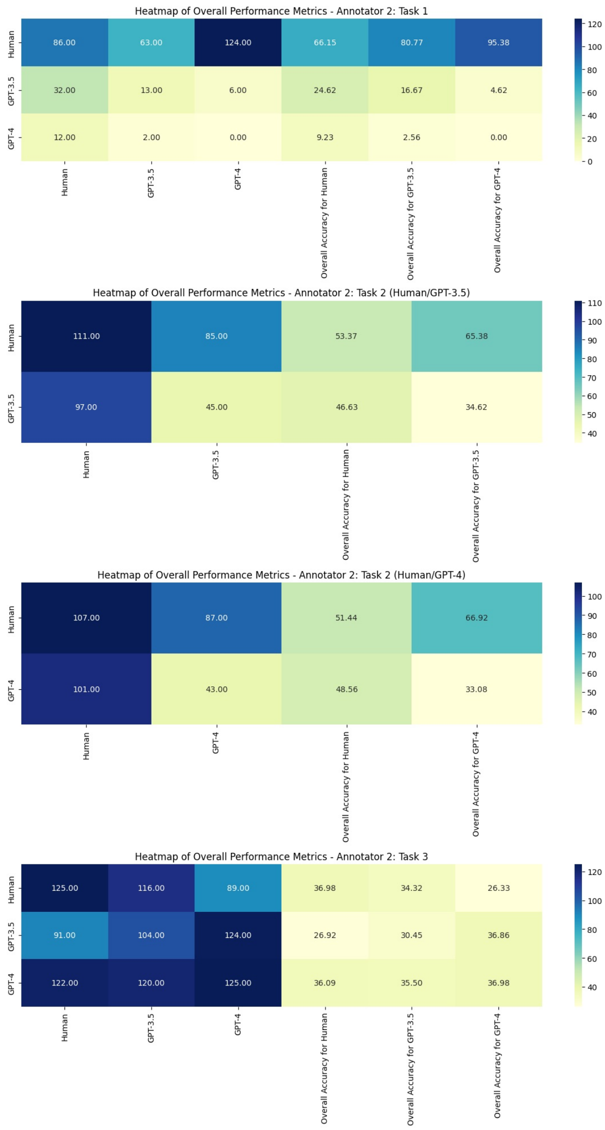

Figure 5: Heatmap of Annotator 2 across Task1, Task2, Task3

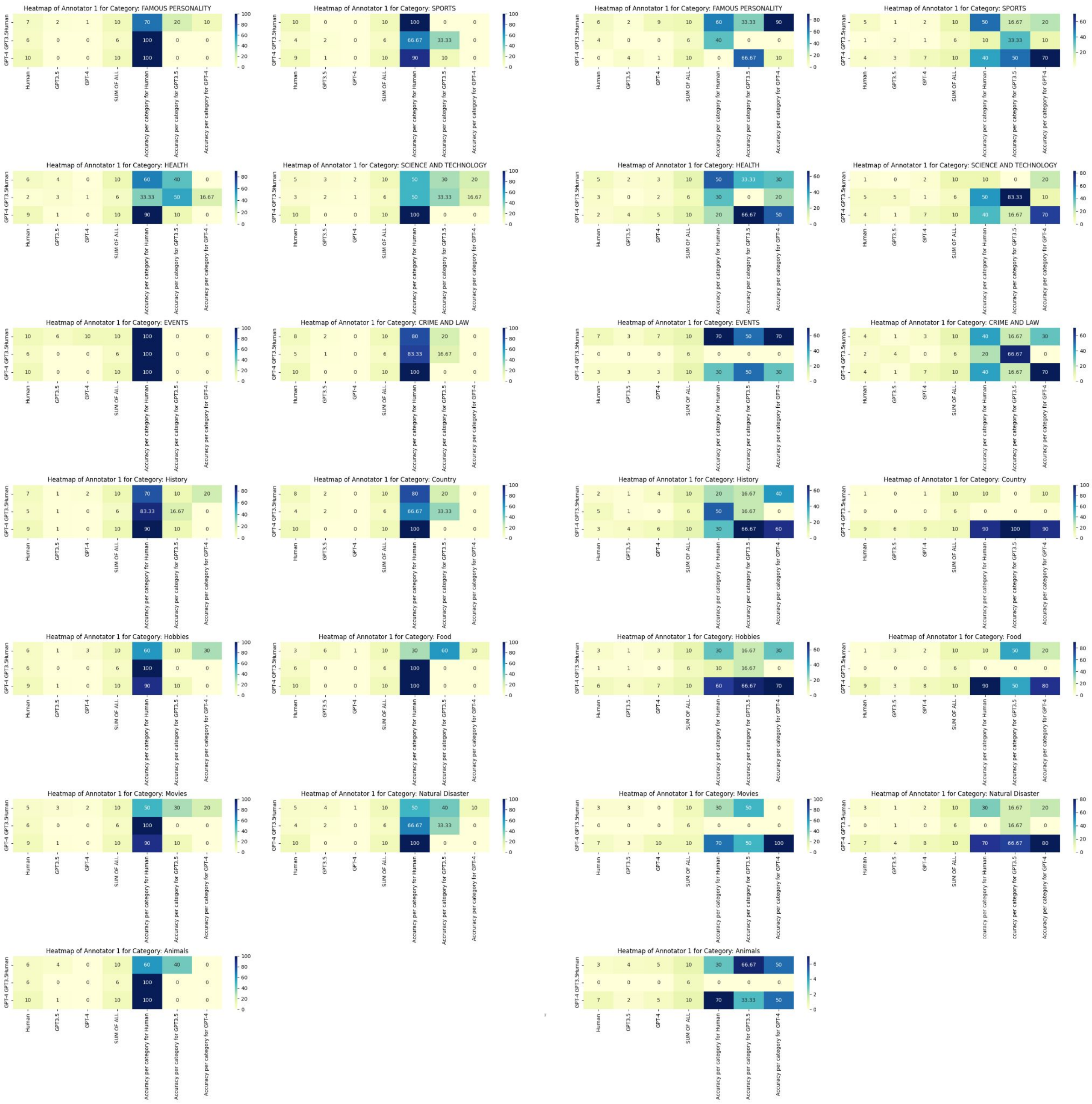

Figure 6: Heatmap Per Category Annotator 1 and 2: Task 1.

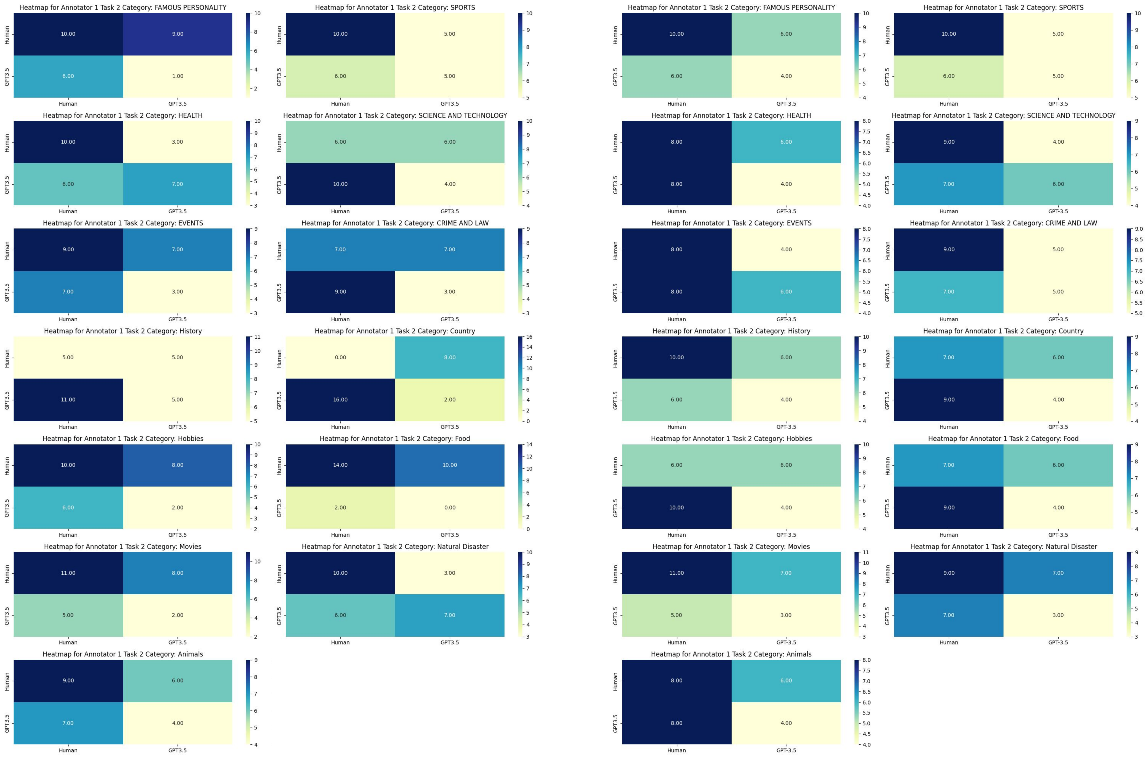

Figure 7: Heatmap Per Category Annotator 1 and 2, Task 2: Human vs GPT-3.5.

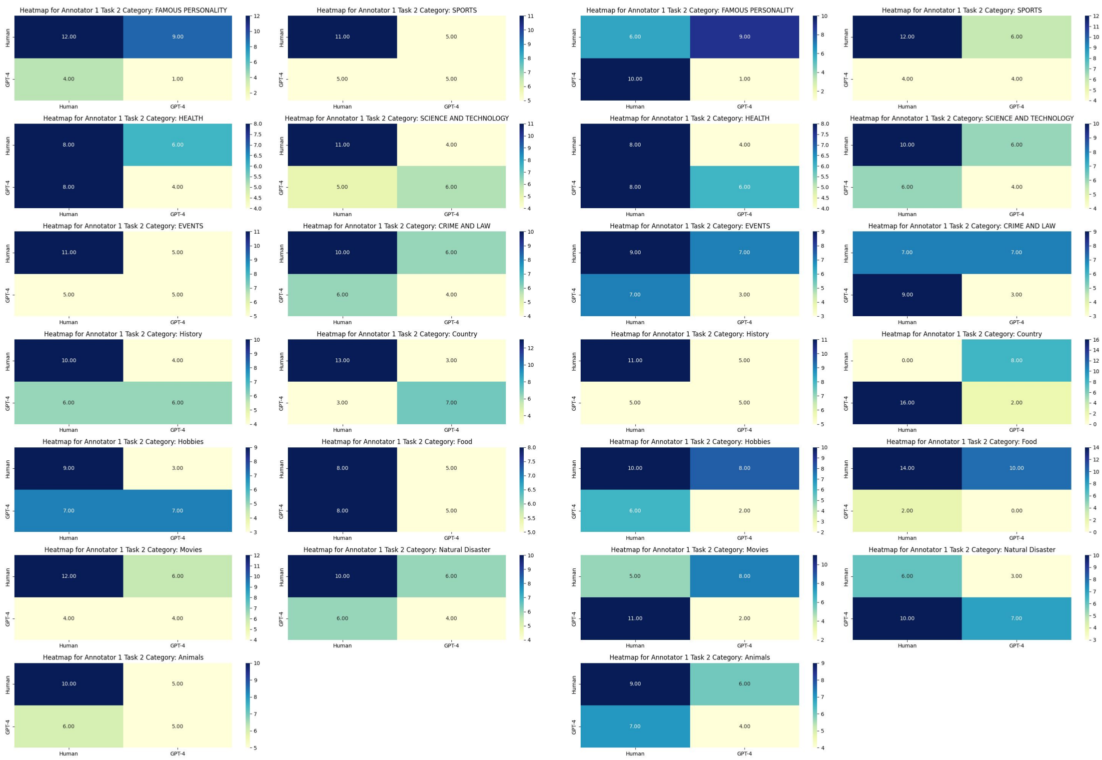

Figure 8: Heatmap Per Category Annotator 1 and 2, Task 2: Human vs GPT-4.

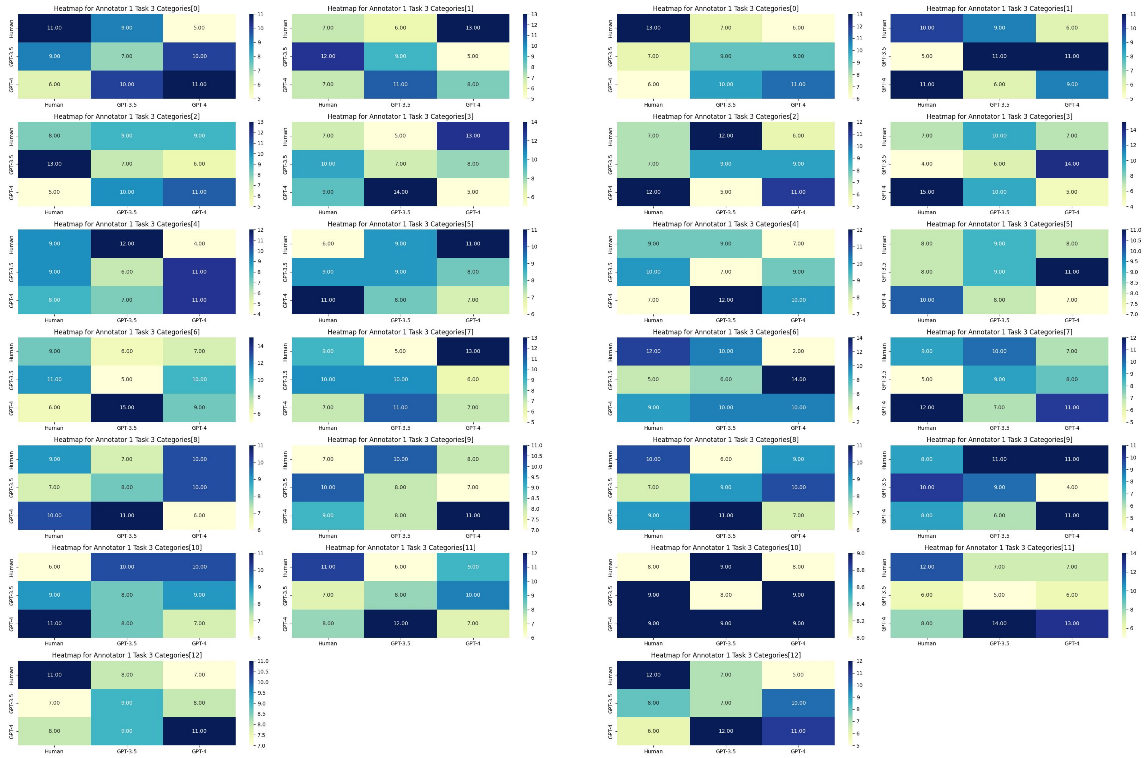

Figure 9: Heatmap Per Category Annotator 1 and 2, Task 3.

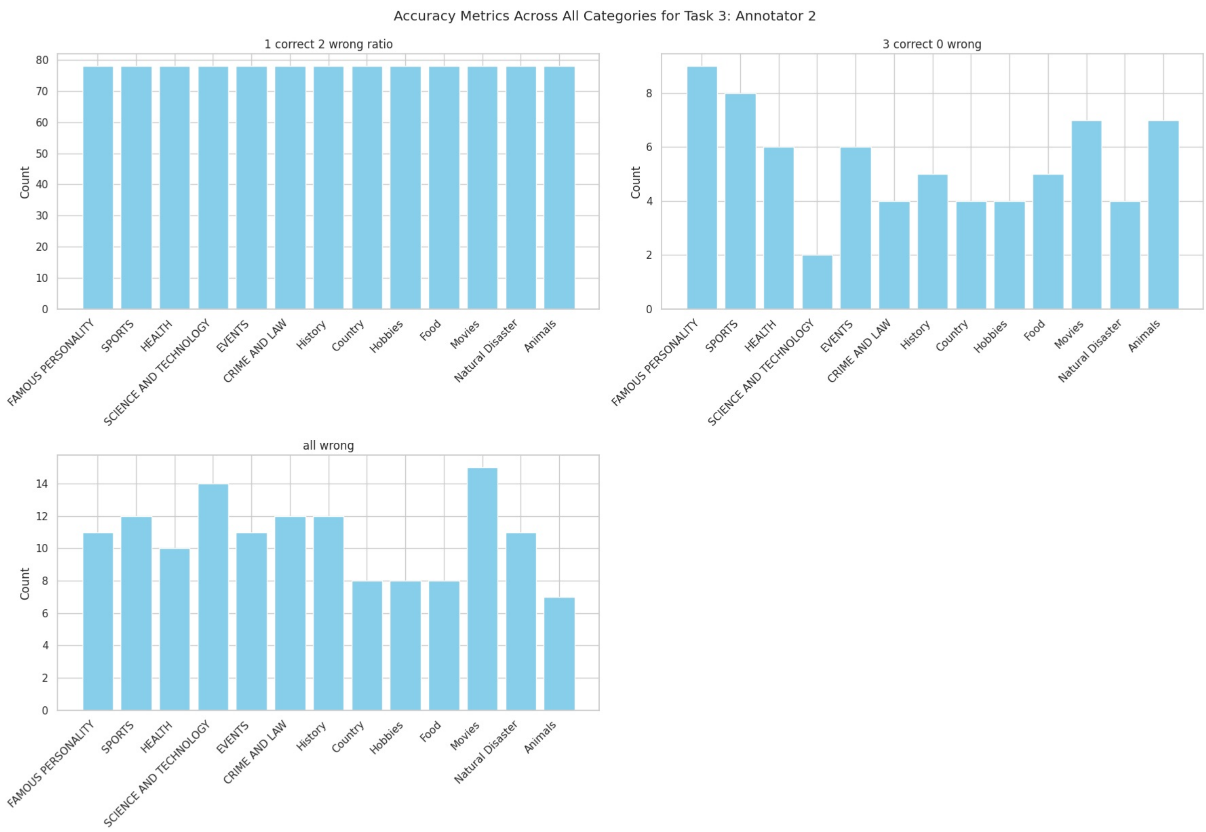

Figure 10: Task 3: OverAll No of Correct Labels,No of Wrong Labels,1 Correct 2 Wrong Labels for Annotator 1.

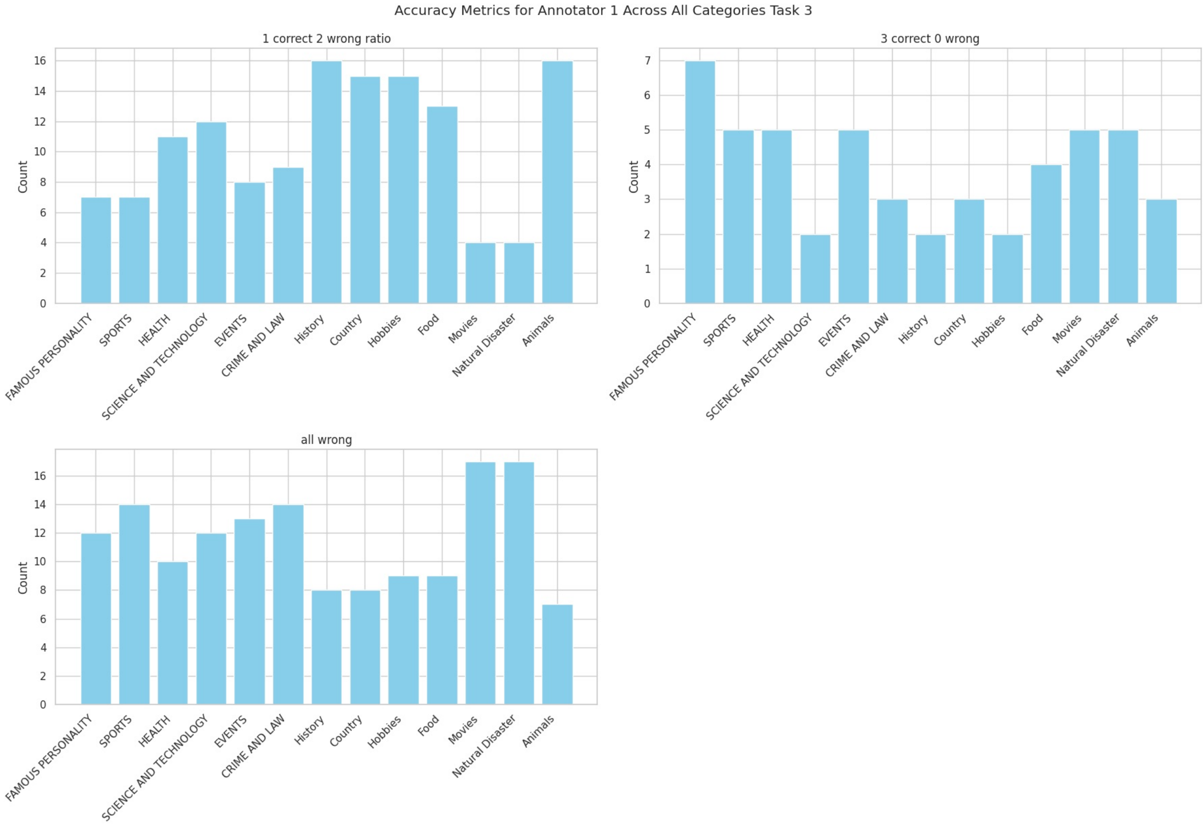

Figure 11: Task 3: OverAll No of Correct Labels,No of Wrong Labels,1 Correct 2 Wrong Labels for Annotator 2.

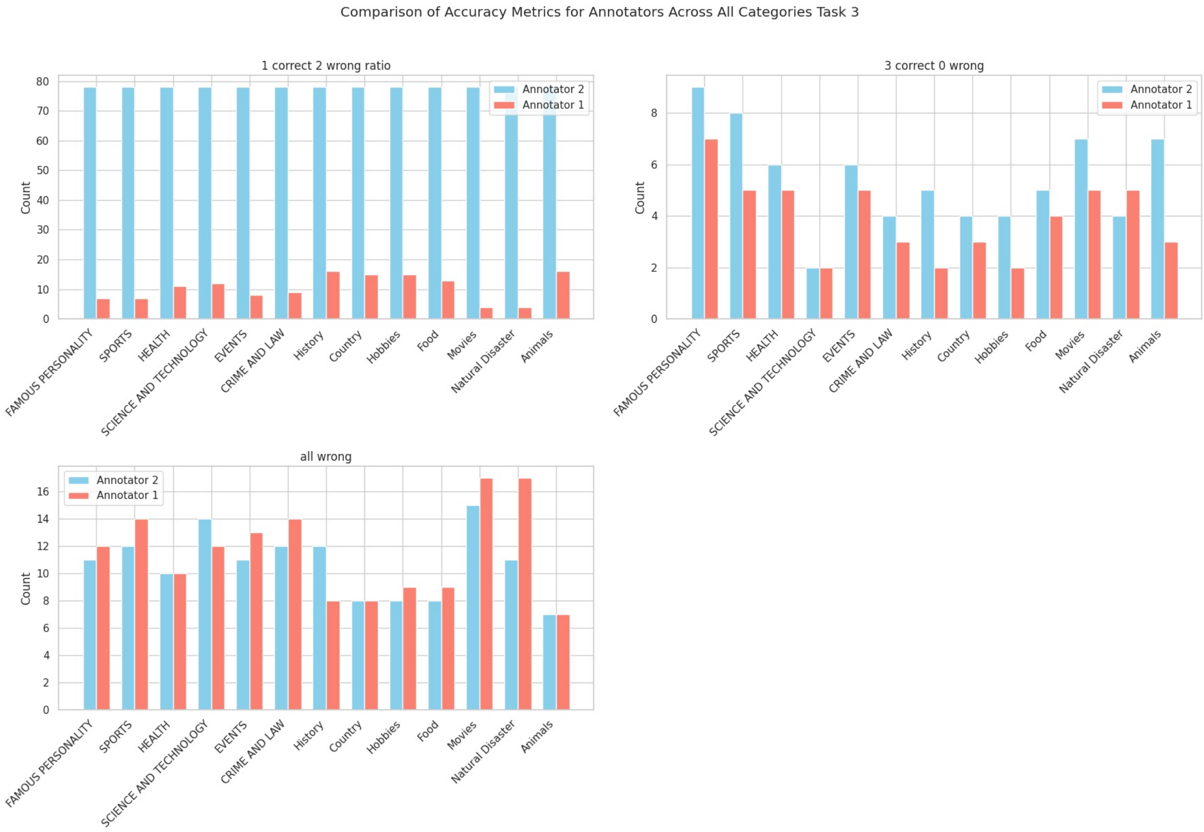

Figure 12: Task 3: Comparison Between OverAll No of Correct Labels,No of Wrong Labels,1 Correct 2 Wrong Labels for Annotator 1 and 2.

Prompt for Analysis
You are a team of AI language analysis experts tasked with determining whether a given sentence was written by a human or generated by a machine. Each of you will analyze the sentence from a unique linguistic perspective and then collaborate to reach a final verdict. The personas are: 1. Linguistic Patterns Persona - Looks for natural flow, idioms, and cultural references common in human writing 2. Word Choice Persona - Analyzes word choice for appropriateness, context, and nuance typical of human writers 3. Grammar Persona - Checks grammatical accuracy including sentence structure, verb conjugation, and subject-verb agreement 4. Consistency Persona - Looks for consistency in tone, style, and context which is often more coherent in human writing 5. Complexity Persona - Evaluates complexity and depth, as human writing often includes subtleties and nuances Here is the sentence to analyze: <sentence> {{SENTENCE}} </sentence> <persona_analysis> Linguistic Patterns Persona: <analysis>...</analysis> Word Choice Persona: <analysis>...</analysis> Grammar Persona: <analysis>...</analysis> Consistency Persona: <analysis>...</analysis> Complexity Persona: <analysis>...</analysis> </persona_analysis> <collaboration> Linguistic Patterns Persona: <thoughts>...</thoughts> Word Choice Persona: <thoughts>...</thoughts> Grammar Persona: <thoughts>...</thoughts> Consistency Persona: <thoughts>...</thoughts> Complexity Persona: <thoughts>...</thoughts> <discussion>...</discussion> <consensus> After analyzing the sentence and discussing our findings, we have reached the following consensus: ... </consensus> <result> Based on our analysis, we classify this sentence as [human written/machine generated]. </result>

Table 15: Prompt template for LLM Evaluation