# Probing the Uniquely Identifiable Linguistic Patterns of Conversational AI Agents

**Iqra Zahid, Tharindu Madusanka, Youcheng Sun** and **Riza Batista-Navarro**
Department of Computer Science, University of Manchester
{iqra.zahid|tharindu.batawalaacharige}@manchester.ac.uk,
{youcheng.sun|riza.batista}@manchester.ac.uk

## Abstract

Considerable effort has been dedicated to detecting machine-generated texts to prevent a situation where the widespread generation of text—with minimal cost and effort— reduces trust in human interaction and factual information online. Our study takes a more refined approach by analysing different Conversational AI Agents (CAAs). By constructing linguistic profiles for each AI agent, the aim is to identify the Uniquely Identifiable Linguistic Patterns (UILPs) for each model and to demonstrate the effectiveness of these UILPs in identifying its respective AI agent using authorship attribution techniques. Promisingly, we are able to classify AI agents based on their original texts with a weighted F1-score of 96.94%. Further, we can attribute AI agents according to their writing style (as specified by prompts), yielding a weighted F1-score of 95.84%, which sets the baseline for this task. By employing principal component analysis (PCA) for dimensionality reduction, we achieve a weighted F1-score ranging from 89.25% to 97.83%, and an overall weighted F1-score of 96.93%.

## 1 Introduction

Recent advances in deep learning and natural language processing have led to the emergence of conversational AI agents (CAAs), hereby referred to as AI agents, which we define as large language models that can generate natural language as a dialogue system. These have been applied in tasks such as question answering (Zhao et al., 2023), fake news detection and abuse detection (Uchendu et al., 2021). The widespread use of AI agents has highlighted the importance of determining the origin of a text (Desaire et al., 2023; Fagni et al., 2021; Mitrović et al., 2023; Fagni et al., 2021; Mitrović et al., 2023; Becker et al., 2023; Islam et al., 2023; Markowitz et al., 2023) and has led to a surge of interest in analysing the linguistic structures within them (Desaire et al., 2023). One noteworthy linguistic aspect that remains unexplored is the determination of whether AI agents possess any uniquely identifiable linguistic patterns (UILPs).

Our research draws inspiration from the linguistic theories of language identity and linguistic patterns within the compositions of individual authors (Nini, 2023; Coulthard, 2004). Specifically, our study undertakes the task of assessing the validity of the aforementioned theories regarding AI agents. As a result, we have meticulously crafted the UILPs for the following five generative large language models: GPT-4[1], GPT-3.5[1], Text-Curie-001[1], PaLM-2[2] and LLaMA2-7b[3], aiming to ascertain the presence of UILPs. The most effective method to confirm the usefulness of an identified UILP is through validation, a process achievable through the task of authorship attribution. Authorship attribution for AI agents is the ability to ascertain the authorship of texts generated by AI agents (Juola, 2008, 2006; Sari, 2018). By establishing authorship, whether proving or disproving, we can reinforce the theory of distinct linguistic patterns in AI agents. We seek to answer crucial questions about the existence of UILPs in AI agents, the linguistic overlap between various text types generated by these models, and the feasibility of identifying AI agents based on their individual UILPs.

The ability to prove the existence of UILPs provides many benefits such as preventing the harmful use of AI agents (e.g., detecting fake news, hate speech, plagiarism). Additionally, this enables the reuse of the UILP in classification tasks, potentially enhancing classification accuracy. We propose a transparent means for linguistic analysis that is more interpretable across different AI agents and forms the central emphasis of this paper.

[1]Model details and source: OpenAI's GPT-3.5. (2021). https://www.openai.com/

[2]Model details and source: Bard: The Language Model for Writing Assistance. (2022). https://www.bardmodel.com/

[3]Model details and source: LLaMA2-7b: A Large Multilingual Language Model for Free-Form Editing. (2023). https://www.llama7b.ai/



*Findings of the Association for Computational Linguistics ACL 2024*, pages 4612–4628
August 11-16, 2024 

Thus far, there has been no investigation on the UILP of AI agents, and there has been only limited comparison of different AI agents and little research indicating if these AI agents can be differentiated from each other based on their linguistic patterns. Moreover, there is a notable absence of analysis of AI agents based on *stylometry*, i.e., the statistical analysis of language often used in the context of forensic linguistics (Rocha et al., 2016). We propose both a feature-based machine learning classification as well as the use of transformer language models for AI agent classification. The research questions (RQs) we aim to answer in this paper are as follows:

RQ1: To what extent can we perform authorship attribution (AA) for AI agents based on their original texts, through the recognition of their UILPs?

RQ2: Can we attribute text to AI agents through the recognition of UILPs in texts that they generated based on different stylistic prompts?

RQ3: How can we measure the linguistic overlap, if any, in outputs from the AI agent when it generates distinct texts?

In addressing the above questions, we have made the following contributions:

- Two new datasets: The first dataset is a collection of original texts created by five AI agents, while the second dataset is an expanded version of the first whereby each text was paraphrased by its respective AI agent according to the following five styles: (a) paraphrased with no specified style, (b) written as a fictitious narrative, (c) written as a tweet, (d) written as a social media blog post and (e) written as an academic article.
- An approach to AI agent attribution based on a Logistic Regression (LR) model trained on linguistic features and a fine-tuned DeBERTa model (He et al., 2021).
- A method for identifying linguistic patterns in the texts generated by the different AI agents based on principal component analysis (PCA).

## 2 Related work

The analysis of authorship attribution encompasses two distinct categories: feature-based and large language model-based classification. Feature-based approaches involve creating a specific feature set for a specific task (Sari, 2018; Juola, 2008). Multivariate linguistic analysis paired with a traditional machine learning classifier is an example of this (Abbasi and Chen, 2008). Approaches using pre-trained transformer language modes have demonstrated superior accuracy with few preprocessing steps (Fabien et al., 2020; Uchendu et al., 2021; Ai et al., 2022). These models significantly outperform traditional models in many cases. Newer approaches use pre-trained transformer models and in some cases, these are combined with linguistic features (Fabien et al., 2020; Sari, 2018).

Posited by Nini (2023), the Principle of Linguistic Individuality states that at any given moment it is exceedingly improbable for two individuals to possess identical linguistic grammars. This principle is aligned with authorship attribution (Coulthard et al., 2016) which assumes that writings from one author would exhibit greater linguistic similarity than writings from a different author (Burrows, 2002; Anthonissen and Petré, 2019). This theory has not been investigated in the case of AI agents, which is what we sought to achieve in our work.

There has been a central focus on GPT models, with an emphasis on distinguishing between text written by humans and those generated by machines using transformer models (Fagni et al., 2021; Mitrović et al., 2023; Solaiman et al., 2019; Uchendu et al., 2021; Bakhtin et al., 2019; Ippolito et al., 2020), or surface-level linguistic features (Desaire et al., 2023; Markowitz et al., 2023) which have been regarded as a limited analysis when studied individually (Schuster et al., 2020). Other studies have utilised a primarily linguistic approach, analysing words and sentiment to distinguish human and machine-generated text (Markowitz et al., 2023). The limitations of previous approaches, compared to the methodology employed in this paper, become evident when considering their emphasis on distinguishing between human and machine-generated content.

These studies lack a comparative analysis of various AI agents and rarely incorporate multivariate stylometric analysis in their evaluation, which would better capture the use of AI agents in generating texts in other scenarios. Munir et al. (2021) investigated the attribution accuracies of synthetic text using transformer models (XLNet) and prior attribution approaches. Other work has shown that traditional authorship attribution approaches cannot fully capture the style of an author when the author is a human (compared to when it is a machine). Machine-generated authorship attribution is a comparatively straightforward task. Machines do not display the same level of linguistic variety inherent in humans limiting their capacity to produce linguistically diverse texts (André et al., 2023). This limitation is imposed by their inability to evolve linguistically without being retrained (Ai et al., 2022). Humans have a wide writing style which means their features and feature usage can differ depending on the text genre (Uchendu et al., 2021). This increases the complexity of authorship attribution tasks.

# 3 Methodology

## 3.1 Model selection

The models used for this project include GPT-3.5, GPT-4, Text-Curie-001 (OpenAI, 2023), PaLM-2 (Anil et al., 2023)[1] and, LlaMA2-7b (Touvron et al., 2023). All of these models are proficient in the natural language generation task with varying levels of sophistication. The Open AI GPT (generative pre-trained transformer)[2] models used in this paper were all trained using reinforcement learning from human feedback (RLHF) on text data, web pages and books, among others. GPT-4 (OpenAI, 2023) is currently the most optimised model; GPT-3.5 has the same capabilities as GPT-4 but operates on a smaller scale. The Text-Curie-001 model is an older, now deprecated model produced by Open AI.

PaLM-2 (Pathways Language Model)[3] developed by Anil et al. (2023) was pre-trained on a large quantity of parallel multilingual corpora, web pages, source code and various other datasets. Proposed by Touvron et al. (2023), LLaMA2-7b (Language Learning and Meaning Acquisition)[4] was trained on textual data using a standard optimiser and RLHF. We refer the reader to Table 1 for details on each model's size (in terms of the number of learned parameters) and the maximum number of tokens in their output.

[1]This model was used via Google's BARD, now known as Gemini (https://gemini.google.com/app)

[2]Introducing GPT models: https://platform.openai.com/docs/guides/gpt

[3]PaLM-2: https://ai.google/discover/palm2/

[4]LLaMA: https://ai.meta.com/blog/large-language-model-llama-meta-ai/

| Model | Creator | Size | # Tokens |
|---|---|---|---|
| GPT-4 | OpenAI | 1.7T | 8192 |
| GPT-3.5 | OpenAI | 175B | 4097 |
| Text-Curie-001 | OpenAI | 6.7B | 2049 |
| PaLM-2 | Google | — | 8192 |
| LLaMA2-7b | Meta | 7B | 2048 |

Table 1: Comparison of AI agents based on their size in terms of the number of parameters (unknown for PaLM-2) and the maximum number of tokens in their output (# Tokens)

## 3.2 Data collection

Data collection was carried out in two phases. In the first phase, a set of 10 prompts was collated, with each prompt corresponding to a news category on the BBC website[5] to cover various topics. The specific topic for each prompt was derived from the headline that was most popular at that time within a particular category. For instance, within the education category, the most prominent headline pertained to the impact of Covid-19 anxieties on academic studies. The topics were selected to ensure a diversity of texts and the provided prompts did not include harmful or sensitive content; therefore, we anticipate the generated text to be devoid of this material. Table 9 in Appendix A provides a list of these prompts. An example of the outputs for the prompts in the different prompt styles can be seen in Table 10 in Appendix B. These prompts were given as input to all the AI agents. Data collection was carried out through two methods: manual input of prompts in the case of PaLM-2 (through BARD), or by utilising APIs in the case of LLaMA2-7b and the GPT models. For each of the 10 prompts, 20 texts were generated. Thus, overall, 200 texts were generated per model. The only exception is PaLM-2, whose generated text corresponds to only nine queries as the model's responses for one of the 10 queries were inadequate, thus leading to the generation of only 180 texts for this model. This dataset will be referred to as our original data. The data was labelled according to the model used, using labels OG0-OG4 (Original-0 to Original-4). We also used only the GPT-generated data (GPT-3.5, GPT-4 and text-curie-001) from the original data in our analysis. This dataset, referred to as the GPT data, was labelled according to the model used using integers from GD0-GD2 (GPT data-0 to GPT data-2) in this dataset.

[5]BBC: https://www.bbc.co.uk/

The second phase pertains to the collection of stylistic data for only GPT 3.5, 4 and Text-Curie-001. We employed only these three AI agents because they responded effectively to the prompt, while other AI agents produced nonsensical or repeated texts. The stylistic data uses the original data to produce paraphrases of this text in different stylistic genres. Firstly, we asked each model to paraphrase the original text in a general manner, i.e., without specifying a specific style. The model was then asked to paraphrase the original text (from the first phase) in four styles: as an academic paper, as a social media post, as a fictitious narrative and as a tweet. These texts were labelled according to the style with labels ranging from S0 to S4 (Style-0 to Style-4) for each stylistic variation in this dataset. For each paraphrasing prompt, 200 texts were generated (corresponding to the original 200 texts generated as part of the first phase). In total, there are 1200 texts for each model: the original 200, a version of those 200 that are general paraphrases and 200 for each of the four above-mentioned styles. This set of data will be referred to as stylistic data. All datasets were split into training and testing sets following a 80:20 partition. No preprocessing steps were applied to the data.[6]

The process of dataset creation posed a challenge, with certain models generating incoherent texts which were variations of the input text, or texts that were too short or too long. This was due to the absence of predefined constraints during the text generation process. The cohesiveness or semantic soundness of texts is not a primary issue in this work as we aim to focus on context-independent linguistic features. Model hallucination was not a significant concern for us, as our work primarily concentrated on extracting linguistic features; hence, the content held minimal importance. However, steps to ensure that generated text was reasonable and free of grammatical errors were taken. As previously mentioned, data collection involved either manually inputting prompts or utilising an API. When employing APIs, texts were generated in small batches of 20-50 rows of text data to guarantee that the model produced coherent text data rather than generating random iterations of a single phrase. Lastly, the final datasets were manually assessed to ensure their suitability for the attribution task. This was assessed by ensuring each row contained enough text (more than 10 tokens), a set number of texts per author (dependent on the dataset), avoided repetitive material and created topic-diverse texts. Diverse prompts were employed to ensure this.

[6]The dataset is available at: https://github.com/iqrazahid05/UILP/

### 3.3 Data evaluation

Data was evaluated using a combination of various automated metrics. We utilised BERTScore and METEOR to assess textual coherence. During data generation, a small subset of the data from the three datasets (original, GPT and stylistic) was assessed by the data collector (the first author of this paper). The results for the automated metrics are presented in Tables 2 and 3. For both metrics, results closer to one indicate higher textual similarity and increased text cohesiveness.

| Meteor Scores | |
|---|---|
| Original | 0.8333 |
| GPT | 0.8331 |
| Stylistic | 0.8315 |

Table 2: METEOR scores for the original, GPT and stylistic datasets

| BERTScore Comparisons | | | |
|---|---|---|---|
| Comparison | Precision | Recall | F1 Score |
| Original vs GPT | 0.8949 | 0.8875 | 0.8911 |
| Original vs Stylistic | 0.8046 | 0.8026 | 0.8030 |
| GPT vs Stylistic | 0.8041 | 0.8059 | 0.8044 |

Table 3: BERTScore comparisons for the original, GPT and stylistic datasets

### 3.4 Writeprints as feature representation

Abbasi and Chen (2008) proposed the Writeprint feature set: a set of linguistic features that ultimately represent the distinctive writing style of an author in an authorship attribution task. The proposed feature set is largely composed of dynamic features, which are context-dependent, an example of which is the presence of certain word, unigrams or bigrams. For example, the presence of the word bigram *"yours sincerely"* could be indicative of a particular author when writing emails. However, the same author is unlikely to use the same bigram in a different context, e.g., when writing an academic article. Thus, to represent an author's writing style regardless of context (or textual genre), we extended the original Writeprint to include static features, which are context-independent and are present in a large percentage of texts irrespective of the genre. The extended feature set differs from the original Writeprints in that the former encompasses previously unexplored aspects of a text, such as phonology, morphological irregularities, ellipsis, and omission. Our Extended Writeprint (EWP) is provided in full in Appendix C. These features were extracted from the texts generated by each of the AI agents of interest with the aid of existing Python packages, e.g., spaCy (Honnibal et al., 2020) and NLTK (Bird, 2006). This results in a unique linguistic profile for each model, which is used in two ways: to determine the most informative features representing the UILP of each of our AI agents of interest (Section 3.5) and to train traditional machine learning-based classification models to attribute a text to its AI agent (Section 3.6).

### 3.5 Analysing the UILP of AI agents

We employed principal component analysis (PCA) to assess the top 100 most informative linguistic features that represent each model (based on its generated texts). We also assess the collective top 100 most informative linguistic features. PCA was performed on the standardised feature counts. Subsequently, we quantified the degree of overlap among these top 100 features across the various models. Instead of necessitating the training or retraining of pre-existing language models for attributing texts generated by AI agents, we advocate for a feature-based approach coupled with a machine learning-based classifier. The advantage of employing a feature-based approach lies in its efficiency, requiring less time and computational resources. By employing a feature-based approach, we can ensure consistent attribution accuracies regardless of when or by whom the text was generated. This is achieved by the ability to identify distinctive linguistic patterns unique to each AI agent.

We identified unique features for each model based on the most informative features identified by PCA. These unique features were then extracted from the writeprint of the texts. Authorship attribution was then performed using these uniquely occurring features.

### 3.6 Classification models

We cast authorship attribution as a multi-class classification problem, whereby a model takes a given text as input and outputs a label that corresponds to any one of the five AI agents.

A variety of traditional machine learning-based models were trained as classifiers. These include Support Vector Machine (SVM), Random Forest (RF) and Logistic Regression (LR) models. Each of these models was trained on the EWP features described in Section 3.4, using optimised parameter values which were defined through the use of grid search. This allowed us to set a baseline and quantify the extent of any performance improvements. We computed the standard deviation (SD) over five runs. Our results show that the SD in all experiments is low, indicating that the performance scores tend to cluster around the mean. This consistency highlights the stability of our results.

Additionally, we aimed to assess the attribution performance in comparison to a transformer-based language model given that transformer models have demonstrated superior performances in classification tasks (Vaswani et al., 2023; Fabien et al., 2020). In this case, we selected the Decoding-enhanced BERT with Disentangled Attention (DeBERTa) model as it has outperformed other transformer models in a variety of classification-related tasks (He et al., 2021). Details of the hyperparameters used in training the machine learning and transformer-based language models can be found in Appendix D and E. All experiments were run on Google Colab using the A100 GPU accelerator. Due to the high computational power required to run the DeBERTa model, the results presented are based on a single run.

Prior approaches tend to overlook the identification of distinctive patterns, opting instead for a multivariate dynamic feature extraction technique. Such techniques are text, author and content specific due to dynamic feature selection (Ai et al., 2022; Sari, 2018). The emphasis here lies in discerning unique patterns that can be utilised to identify the AI agents of interest, regardless of the text they produce, with consistent results.

## 4 Evaluation Results and Discussion

### 4.1 Attribution of original Texts

Table 4 presents the results for authorship attribution based on the original data. The EWP features were extracted from all the texts and the methodology was applied. From the results, we can see that the optimised DeBERTa model obtained the highest weighted F1-score at 99.11%. However, it is worth noting that the discrepancy in F1-scores across all models is at most merely 5.23% demonstrating competitive performance across all models. When the extended feature set is combined with

| Data | Model | Accuracy | W-F1 | SD |
|---|---|---|---|---|
| Original Data | SVM | 93.87 | 93.88 | 0.00 |
| | RF | 96.54 | 96.54 | 0.37 |
| | LR | 96.94 | 96.94 | 0.00 |
| | DeBERTa | 99.11 | 99.11 | - |
| GPT Models | SVM | 94.11 | 94.17 | 0.00 |
| | RF | 96.67 | 96.67 | 0.00 |
| | LR | 97.50 | 97.50 | 0.19 |
| | DeBERTa | 99.11 | 99.11 | - |
| Stylistic Data | SVM | 95.56 | 95.56 | 0.00 |
| | RF | 95.25 | 95.24 | 0.25 |
| | LR | 95.83 | 95.84 | 0.00 |
| | DeBERTa | 88.00 | 88.00 | - |

Table 4: Performance metrics for different data groupings. The accuracy, weighted F1-score (W-F1) and standard deviation (SD) when using optimised SVM, LR, RF and DeBERTa models. Standard deviation is calculated after 5 runs.

an optimised ML classifier, the weighted F1-score ranges from 93.88% to 96.94%. This demonstrates the existence of UILPs in each AI agent due to the attribution success of each model displaying a minimum weighted F1-score of 93.88%.

From the results in Table 4, we can see that DeBERTa has the highest weighted F1-score at 99.11%. In this experiment, the discrepancy in F1-scores across all models is 4.94%. Since all the compared models are OpenAI-engineered, it is reasonable to anticipate that they exhibit similar linguistic patterns in their generated texts hence the lower F1-scores across all experiments. This model displays an impressively competitive performance, with the optimised LR model having a weighted F1-score of 97.50%, which is only a 1.61% drop when compared to a fine-tuned DeBERTa model.

### 4.2 Attribution of stylistic Texts

We perform cross-genre authorship attribution by investigating the attribution of stylistic texts. We examine the attribution performance for all AI agents on different stylistic data.

The results of stylistic attribution for GPT models are presented in Table 4. As aforementioned, since all models are OpenAI-engineered we expect some linguistic commonalities across different genres of text. Here we attempt to attribute all texts (original, paraphrase, social media posts, tweets, academic articles and fictitious narratives) to their respective AI agent. The results support the notion of models demonstrating a UILP in their generated stylistic texts as well as the notions posited by Juola (2008); Sari (2018); Coulthard (2004) who suggested that these UILPs can be identified across different textual genres. Notably, lower results can be expected when performing cross-genre attribution. This accounts for the 11.11% reduction in the weighted F1-score when comparing the original data to the stylistic data using optimised DeBERTa models. We observe a deduction of 1.1% in the weighted F1-score when using an optimised LR model and a 19.62% deduction when comparing the classification abilities of the default LR model on the original data and the stylistic data. To conclude, irrespective of the text's style, each AI agent can be identified according to the highest weighted F1-score attained which stands at 95.84%. This reaffirms the notion that classification performances decrease across genres due to varying linguistic patterns (Stamatatos, 2016).

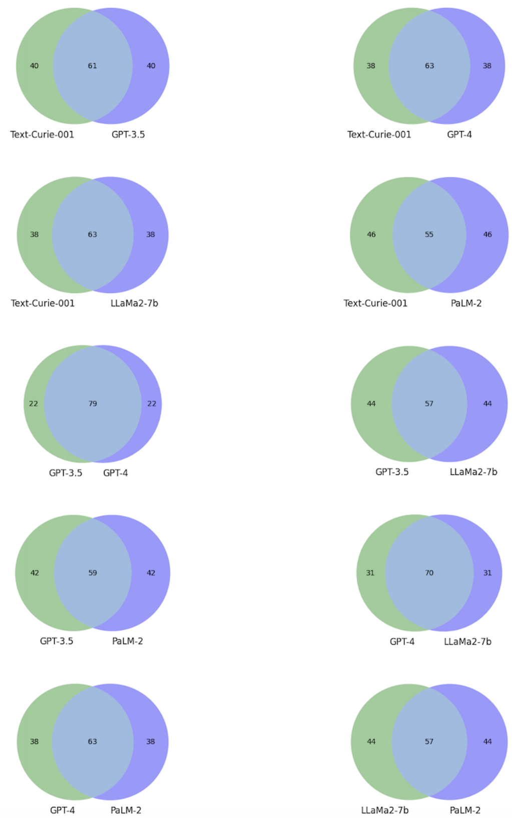



Figure 1: Overlap for the top 100 most informative linguistic features extracted based on our EWP using PCA for all AI agents.

### 4.3 Attribution based on PCA of features

In this section, we identify the top 100 most informative linguistic features across all AI agents and the top 100 most informative linguistic features for each AI agent. We then assess the extent to which attribution can be performed based on these.

| Accuracy and Weighted F1-Score | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| AI agent | GPT-3.5 | | GPT-4 | | LLaMA2-7b | | PaLM-2 | | Text-Curie-001 | |
| GPT-3.5 | 80.52 | 80.49 | 82.50 | 82.50 | 78.06 | 78.05 | 88.89 | 88.89 | 90.85 | 90.84 |
| GPT-4 | 78.16 | 78.16 | 87.50 | 87.50 | 72.95 | 72.94 | 83.54 | 83.54 | 90.91 | 90.91 |
| LLaMA2-7b | 65.64 | 65.63 | 77.16 | 77.14 | 66.67 | 66.67 | 75.00 | 75.04 | 94.75 | 94.74 |
| PaLM-2 | 82.05 | 82.05 | 84.67 | 84.62 | 86.42 | 86.43 | 79.49 | 79.49 | 97.31 | 97.30 |
| Text-Curie-001 | 98.77 | 98.77 | 95.24 | 95.24 | 98.77 | 98.77 | 97.56 | 97.56 | 98.79 | 98.77 |
| Overall | 81.63 | 81.00 | 85.71 | 85.42 | 81.12 | 80.45 | 85.01 | 85.36 | 94.39 | 94.38 |

Table 5: Accuracy and weighted F1-scores for models based on their top 100 most informative linguistic features extracted from the EWP using PCA analysis. Attribution was performed for each model and then for the entire original dataset using an optimised Logistic Regression model

For all the original data, we extracted our EWP features. Subsequently, we conducted PCA to identify the top 100 most informative linguistic features across the entire dataset. Attribution was carried out using these selected top 100 features; the accuracy of each model was then computed and provided in Table 5. When performing attribution using only the top 100 most informative linguistic features, we found that Text-Curie-001 has the highest weighted F1-score for any model and has a self-identifying weighted F1-score of 98.77%. LLaMA2-7b obtained the lowest performance, with a weighted F1-score of 66.67% when identified using its individual top 100 feature set.

These results support the theory of linguistic individuality (Nini, 2023) as the AI agents do not have identical grammars even though the training material, methods and developers are the same or similar. This can be seen explicitly in the analysis of the Open AI GPT models, whereby the F1-score varies from 96.93% to 88.25%, showing a slight discrepancy of 8.68%. Each AI agent struggles to distinguish itself when using its own top 100 most informative features. However, we found that this is due to the substantial overlap in these features, as demonstrated in Figure 1. On average, they share more than 50% of their top 100 features with another AI agent. This clarifies why, in Table 5, we observe an absence of a distinct pattern in AI agents' ability to identify themselves through their top 100 features. There are noticeable instances of misclassification concerning GPT-3.5 and GPT-4. The relatively poorer attribution of GPT-3.5 and GPT-4 can be explained by the fact that both models are OpenAI-engineered and have similar training processes.

Further investigation was performed to determine if AI agents can be identified based on their unique feature sets. We conducted a comparison of the top 100 features across all AI agents and identified features unique to each model (see Tables 6 and 7). Based on the analysis of the top 100 most information features across all AI agents, our results reveal the ability to attribute each AI agent with weighted F1-scores ranging from 89.25% to 97.83%. These results suggest the existence of several linguistic commonalities. However, as seen in Table 6 classification results improve when investigating unique features per model as opposed to a *"one size fits all"* approach (Uchendu et al., 2021).

After obtaining the set of distinctive features for each model, we moved on to the original dataset containing approximately 300 features. For each model, we exclusively extracted the features that were unique to that model. For example, during the attribution for GPT-4, we isolated features that were uniquely associated with GPT-4 in its top 100 most informative features. GPT models exhibited greater morphological diversity among

| AI agent | Accuracy | W-F1 | SD |
|---|---|---|---|
| GPT-3.5 | 91.60 | 89.25 | 0.03 |
| GPT-4 | 97.63 | 95.50 | 0.01 |
| LLaMA2-7b | 100 | 97.83 | 0.00 |
| PaLM-2 | 95.35 | 93.17 | 0.01 |
| Text-Curie-001 | 100 | 96.97 | 0.00 |
| All | 96.93 | 96.93 | 0.02 |

Table 6: Accuracy and weighted F1-score (W-F1) for each AI agent when performing authorship attribution using only their unique features

| AI agent | Accuracy | W-F1 | SD |
|---|---|---|---|
| GPT-3.5 | 86.42 | 86.17 | 0.00 |
| GPT-4 | 86.08 | 87.18 | 0.00 |
| LLaMA2-7b | 93.34 | 100 | 0.02 |
| PaLM-2 | 94.74 | 90.00 | 0.00 |
| Text-Curie-001 | 98.77 | 97.56 | 0.00 |
| All | 91.84 | 91.81 | 0.00 |

Table 7: Accuracy and weighted F1-score (W-F1) of attribution using an optimised LR model trained on the top 100 most informative linguistic features extracted using PCA across all datasets

these unique features compared to LLaMa7-2b and PaLM-2. In contrast, the unique feature sets of LLaMa7-2b and PaLM-2 predominantly included function words. These specific features were then extracted for every model from the comprehensive set of 300 features. Subsequently, we performed attribution analyses for each model based on this refined set of features. For example, we identified and extracted all features that were uniquely identified in the top 100 features. We then extracted GPT-4's unique features for all other AI agents and attempted attribution using this unique feature set. The differences in results were significant: the weighted F1-scores ranged from 86.17% to 100% when using optimised hyperparameters. The results support the theory that when investigating an AI agent's inherently unique features, one can attribute each AI agent with greater success. Further results on the attribution success for each model can be seen in Table 6.

The subsequent phase involved conducting PCA for each model and extracting the most informative top 100 features. Following this, we attempted the attribution for all models using these top 100 features. The results, shown in Table 5, indicate that only LLaMA2-7b could successfully self-identify as the most similar AI agent based on these features. A more in-depth linguistic examination of these features revealed that PCA features are predominantly comprised of static features, defined as context-independent and frequently occurring attributes. Furthermore, the diagrams in Figure 1 illustrate substantial feature overlap among different models when analysing 300 features. This supports the theory of Linguistic Uniqueness (Nini, 2023) and the existence of a UILP as it is evident that each model has a set of features that it does not share with the others. These results pertain solely to the original data, with accuracies and weighted F1-scores obtained using the LR algorithm.

### 4.4 Linguistic analysis

For each model, rather than extracting all features specified in the EWP, we reduced the feature set to include only linguistic features associated with each specific linguistic category (details of the features and their categories are provided in Table 11 in Appendix C). Attribution was subsequently conducted for the original data using these refined feature sets. The results of this classification are presented in Table 8.

Individual accuracy scores for each linguistic category and the overall dataset were computed. Tagging and *n*-gram categories achieved the highest weighted F1-score among all ML classifiers. This can be attributed to several factors. Firstly, the presence of over 100 different part-of-speech and dependency tags as well as *n*-grams adds a significant level of linguistic diversity to the dataset. This category also encompassed the labels for different sentence types e.g. the count of passive sentence constructions. Furthermore, research has established that AI agents employ repetitive sentence structures to maintain cohesiveness, and this makes tags a particularly identifiable linguistic structure (Mitrović et al., 2023; Markowitz et al., 2023). Forensic research has also continually highlighted *n*-grams as an extremely identifiable linguistic feature in authorship attribution (Sari, 2018). It is still important to note that there is greater variability in the weighted F1-scores with the highest F1-score for any classifier being 91.79% (for RF) and the lowest at 73.30% (for LR) creating a difference of 18.54% between classifiers.

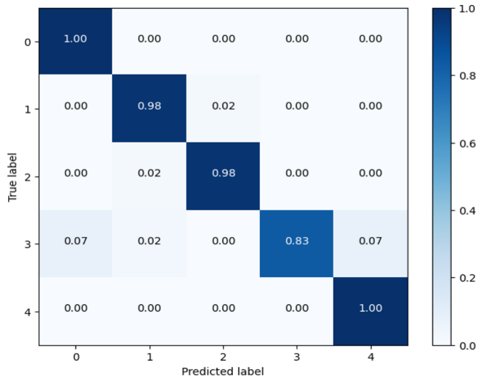
| | Accuracy and Weighted-F1 | | | | | |
|---|---|---|---|---|---|---|
| | SVM | | RF | | LR | |
| Word lists | 89.80 | 89.73 | 88.72 | 88.30 | 84.18 | 84.31 |
| Symbols | 84.18 | 83.70 | 91.33 | 91.26 | 78.57 | 78.34 |
| Tags | 87.76 | 87.66 | 91.33 | 91.34 | 89.80 | 89.77 |
| Syntax | 75.51 | 76.21 | 77.04 | 77.37 | 72.96 | 73.30 |
| Semantic | 79.08 | 79.06 | 84.69 | 84.76 | 79.59 | 79.38 |
| Lexical | 90.51 | 90.33 | 91.33 | 91.33 | 89.80 | 89.86 |
| N-gram | 91.84 | 91.79 | 90.31 | 90.31 | 90.31 | 90.18 |

Table 8: Accuracy and Weighted F1-scores for individual linguistic categories on the original data

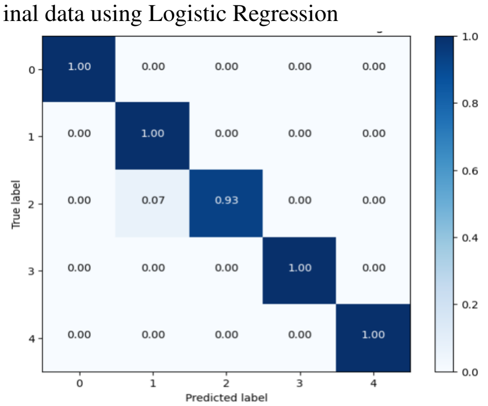
In comparing the linguistic features of two texts from two AI agents in a qualitative assessment, distinctive patterns emerge, suggesting potential variations in author style and expression. Text T1 refers to the text generated by PaLM-2 and Text T2 is the one from GPT-3.5 as seen in Table 16 in Appendix F. Both texts share some features, indicating commonalities in sentence structures and grammatical constructions. Despite sharing five common POS and dependency tags, both texts display between 8 to 21 unique dependency and POS tags, signifying a common syntactic foundation with specific linguistic constructions that differentiate their styles. Notably, T1 employs comparative adjectives, while T2 includes modal verbs, showcasing distinctive choices that may reflect variations in tone or style. In the realm of authorship attribution, these linguistic differences underscore the potential for the texts to be perceived as the work of different authors, as individual writing habits and preferences become apparent through their unique linguistic patterns. See Tables 17 and 18 in Appendix F for further details on the feature counts and for the subset of features extracted for this assessment.

### 4.5 Error analysis

The error analysis was conducted on the original dataset. This can be attributed to its incorporation of all AI agents and utilisation of the EWP. Further, attribution was performed using both an ML and transformer model (in this case Logistic Regression DeBERTa (He et al., 2021)).




(a) Confusion matrix for the attribution of the Original data using Logistic Regression




(b) Confusion matrix for the attribution of the Original data using DeBERTa

Figure 2: Key: 0 = Text-Curie-001; 1 = GPT-3.5; 2 = GPT-4; 3 = LLaMa2-7b; 4 = PaLM-2

In Figure 2a and 2b we see the classification outputs from LR and DeBERTa. The DeBERTa model exhibited a total of three misclassifications. All three instances involved GPT-4 data being incorrectly labelled as GPT-3.5. The explanation for this lies in the fact that both models undergo the same training process. Both are OpenAI authored and additionally, GPT-3.5 is the predecessor of GPT-4. The LR model displayed a total of 9 misclassifications. There is one instance of GPT-4 misclassification as GPT-3.5, a mistake made by DeBERTa. All other misclassifications were of LLaMa2-7b; this AI agent was incorrectly classified as Text-Curie-001, GPT-3.5 and PaLM-2. Based on a linguistic assessment of the misclassified data, we see that the instances of misclassified LLaMA2-7b data exhibited stylistic variations. These texts tended to be longer on average and had more morphological variation which explains the misclassifications as Text-Curie-001 and GPT-3.5. Nevertheless, both models exhibited a minimum number of errors, leading us to consider them insignificant. Further fine-tuning and conducting additional linguistic analysis could help mitigate these misclassifications.

## 5 Conclusion and future work

In our study, we have confirmed the presence of Uniquely Identifiable Linguistic Patterns (UILPs) in conversational AI agents. This is supported by high accuracy in attribution for both original and stylistic data, with weighted F1-scores ranging from 93.88% to 96.96% when utilising the Extended Writeprint (EWP) and traditional machine learning-based classifiers. We also demonstrate similar performance when using a fine-tuned DeBERTa model, achieving a 99.11% weighted F1-score. Our results demonstrate that traditional machine learning-based models can obtain competitive attribution performance compared to a fine-tuned DeBERTa model when utilising the EWP for classification. Through PCA analysis, we explored the attribution of AI agents based on their UILPs. Our results show that the combination of our EWP and RF classification effectively supports cross-genre attribution, with weighted F1-scores ranging from 94.17% to 97.50% for the stylistic data. This affirms the principle of linguistic individuality in AI agents, showcasing their UILPs. These findings offer valuable insights into the distinctive linguistic patterns in text generated by AI agents, with potential applications in digital forensics, detecting fake news and plagiarism. Future work will improve both the datasets introduced in this paper by expanding the size and scope of the stylistic prompts. We seek to perform a fine-grained linguistic analysis of a larger set of AI agents cross-lingually.

## Limitations

In our study, text generation using various APIs that make our AI agents of interest accessible proved to be a time-intensive process, limiting the volume of prompts that could be supplied and thus the text that can be generated. Additionally, certain models imposed output constraints. For instance, in the case of PaLM-2, we resorted to manually inputting prompts into BARD due to the unavailability of the API, which was a time-consuming endeavour. Furthermore, some AI agent outputs did not produce cohesive texts (in the case of LLaMA2-7b), or produced very short texts (in the case of Text-Curie-001). Also, only a set of three text genres were investigated: academic articles, fictitious narratives and, tweets and social media posts (the latter two falling under the same genre). To perform cross-genre authorship attribution we must expand this scope to cover a wider array of genres as well as investigate at different levels of formality. Lastly, a study into misclassified instances must be conducted to identify patterns or determine if there is a specific type of error being made by the model.

## Ethics Statement

For this study, the data was sourced from various AI agents, and human involvement was not required. The dataset does not contain any harmful or sensitive content. As there was no human participation and no collection of personal data, an ethics review was not necessary.

## References


Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.*, 26(2):7:1–7:29.

Bo Ai, Yuchen Wang, Yugin Tan, and Samson Tan. 2022. Whodunit? Learning to Contrast for Authorship Attribution.

Christopher M. J. André, Helene F. L. Eriksen, Emil J. Jakobsen, Luca C. B. Mingolla, and Nicolai B. Thomsen. 2023. Detecting AI Authorship: Analyzing Descriptive Features for AI Detection. In *NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence*, Rome, Italy. Copenhagen Business School, Frederiksberg, Denmark.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. PaLM 2 Technical Report.

Lynn Anthonissen and Peter Petré. 2019. Grammaticalization and the linguistic individual: new avenues in lifespan research. *Linguistics Vanguard*, 5(s2):20180037.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or Fake? Learning to Discriminate Machine from Human Generated Text. *arXiv*.

Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2023. Paraphrase Detection: Human vs. Machine Content.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

John F. Burrows. 2002. 'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship. *Lit. Linguistic Comput.*, 17:267–287.

Malcolm Coulthard. 2004. Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, 25(4).

Malcolm Coulthard, Alison Johnson, and David Wright. 2016. *An introduction to Forensic Linguistics: Language in evidence*. Routledge.

Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David Hua. 2023. ChatGPT or academic scientist? Distinguishing authorship with over 99off-the-shelf machine learning tools.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. BertAA : BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. TweepFake: About detecting deepfake tweets. *PLOS ONE*, 16(5):e0251415.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with Disentangled Attention.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822. Association for Computational Linguistics.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023. Distinguishing Human Generated Text From ChatGPT Generated Text Using Machine Learning.

Patrick Juola. 2006. Authorship Attribution. *Found. Trends Inf. Retr.*, 1(3):233–334.

Patrick Juola. 2008. Measuring Linguistic Complexity. *Foundations and Trends® in Information Retrieval*, 1:233–334.

David M Markowitz, Jeffrey Hancock, and Jeremy Bailenson. 2023. Linguistic Markers of Inherently False AI Communication and Intentionally False Human Communication: Evidence from Hotel Reviews.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. *ArXiv*.

Shaoor Munir, Brishna Batool, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2021. Through the looking glass: Learning to attribute synthetic text generated by language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1811–1822, Online. Association for Computational Linguistics.

Andrea Nini. 2023. *A Theory of Linguistic Individuality for Authorship Analysis*. Elements in Forensic Linguistics. Cambridge University Press.

OpenAI. 2023. GPT-4 Technical Report.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. *IEEE transactions on information forensics and security*, 12(1):5–33.

Yunita Sari. 2018. *Neural and non-neural approaches to authorship attribution*. Ph.D. thesis, University of Sheffield, UK. British Library, EThOS.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. 2020. The Limitations of Stylometry for Detecting Machine-Generated Fake News. *Computational Linguistics*, 46(2):499–510.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models.

Efstathios Stamatatos. 2016. Authorship Verification: A Review of Recent Advances. *Research on computing science*, 123:9–25.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A benchmark environment for Turing test in the age of neural text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2001–2016, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models.

## Appendix A Prompts for AI agents

| | News category | Prompt |
|---|---|---|
| 1 | Cost of living | *Write me an essay on rising house prices in 2023* |
| 2 | Climate | *Write me an essay on what the UK can do to reduce carbon emissions* |
| 3 | Tech | *Write me an essay on Facebook's transfer of European data to US servers* |
| 4 | Politics | *Write me an essay on the UKs ministerial code* |
| 5 | Culture | *Write me an essay on the Glastonbury festival in the UK* |
| 6 | Science | *Write me an essay on the womb condition adenomyosis* |
| 7 | Family and Education | *Write me an essay on how Covid anxiety is affecting students and exams* |
| 8 | War in Ukraine | *Write me an essay on Zelensky's contribution in the war in Ukraine* |
| 9 | UK | *Write me an essay on the Madeleine McCann police search in Portugal* |
| 10 | World | *Write me an essay on the Star Wars original studios* |

Table 9: The prompts used to collect AI agent-generated texts. All original texts were produced all by AI agents: GPT-3.5, GPT-4, Text-Curie-001, PaLM-2 and LLaMA2-7b. In contrast, only GPT-3.5, GPT-4 and Text-Curie-001 were used in generating texts according to stylistic variations (paraphrase, as a social media post, as a tweet, as an academic article and as a fictitious narrative

## Appendix B Data Examples

| Style | Text |
|---|---|
| Original | In 2023, a pervasive and pressing issue has gripped households across the globe — the cost of living crisis. As economic landscapes evolve and global conditions shift, individuals and families find themselves grappling with the ever-increasing burden of expenses. |
| Paraphrase | In 2023, households worldwide are confronted with a widespread and urgent problem — the crisis of rising living costs. As economic environments undergo changes and global circumstances shift, individuals and families are contending with a growing weight of expenditures. |
| Social media post | Feeling the pinch of the 2023 cost of living crisis! It's a tough reality we're all navigating together. From skyrocketing prices to tightening budgets, the struggle is real. Let's open up the conversation and share our tips on stretching those hard-earned pounds, supporting one another through these challenging times. #CostOfLivingCrisis #CommunitySupport #BudgetingTips #TogetherWeCanHandleIt |
| Tweet | The cost of living is soaring in 2023, and it's hitting hard! From groceries to gas, it's becoming more challenging to make ends meet. Let's keep the conversation going on how we can tackle this #CostOfLivingCrisis together and support those facing financial strain. #Economy2023 |
| Academic Paper | The year 2023 has witnessed a growing concern worldwide regarding the cost of living. The term "cost of living" encompasses the expenses individuals and families incur to maintain a reasonable standard of living, including housing, food, transportation, healthcare, education, and other essentials. This essay seeks to provide a comprehensive analysis of the cost of living crisis in 2023, focusing on its underlying causes, economic implications, and potential policy measures to mitigate its effects. |
| Fictitious narrative | In the year 2023, as the calendar pages turned, people across the nation found themselves entangled in a relentless and unforgiving cost of living crisis. The once-stable balance of life, as they knew it, had been upended, and every aspect of their daily existence was impacted. |

Table 10: The GPT-3.5 output for the prompt *"Write me a <stylistic_text> on the cost of living crisis in 2023"*, where <*stylistic_text*> is replaced by one of paraphrase, social media post, tweet, academic article and fictitious narrative

## Appendix C The Extended WritePrint

| Category | Feature | Description |
|---|---|---|
| Lexical | Token-based | Word length |
| | | Sentence length |
| | | Average sentence count, Average word count |
| | Character-based | Upper- and lower-case distribution |
| | | Digit frequency |
| | Word length distribution | One to ten plus letters |
| | Top n-grams | Top 50 occurring tri and bi grams |
| | Special characters/punctuation | Frequency counts |
| | Vocabulary richness | Type-token ration (TTR) |
| | | Text repetitiveness rate (TRR) |
| | Hapax Legomena | Frequency counts |
| | Clipping | Process of shortening words at any word boundary: e.g., “Advertisement” to “Ad” |
| Syntactic | Tagging | Part-of-Speech (POS) tags |
| | | Dependency tags |
| | Term replacement/omission | Ellipsis: e.g. [full sentence] “I like coffee and she likes tea” to [elliptical sentence] “I like coffee, and she” |
| | | Substitutions: e.g. [full sentence] “John went to the store. John bought back milk” to [substituted sentence] “John went to the store. He bought back milk” |
| | Morphological Variation | Irregular patterns: |
| | | - Present participle form |
| | | - Plural forms |
| | | - Past tense form |
| | | - Past participle form |
| | | - Plural form (-ies, -ves, es) |
| | | - Possessive form |
| | | - Comparative and Superlative form |
| | | - Singular form (-y, -o) |
| | Sentence types | Simple, Complex, Compound |
| | | Declarative, Interrogative, Exclamatory, |
| | | Imperative, Conditional, Comparative, Passive |
| Semantic | Sentiment scores | |
| | Synonym/Homonym counts | |
| Other | Phonetic | Alliteration |
| | | Assonance |
| | | Consonance |
| | Word lists | Function words |
| | | Acronyms/Slang |

Table 11: The Extended WritePrint (EWP). This feature set consists of static (context-independent) and dynamic (context-dependent) features

## Appendix D Hyperparameter settings for the DeBERTa model

| Hyperparameter | Amended value |
|---|---|
| num_train_epochs | 6 |
| train_batch_size | 16 |
| eval_batch_size | 16 |
| gradient_accumulation_steps | 4 |
| n_gpu | -1 |
| max_seq_length | 512 |
| class_weight | Custom labels specified |
| early_stopping_patience | 2 |
| early_stopping_delta | 0.01 |

Table 12: The hyperparameters used in training the DeBERTa model (He et al., 2021)

## Appendix E Hyperparameter settings for the traditional machine learning-based classification models

| Hyperparameter | Amended value |
|---|---|
| max_depth | None |
| min_samples_leaf | 1 |
| min_samples_split | 5 |
| n_estimators | 300 |
| class_weights | Balanced |

Table 13: The hyperparameters used in training the Random Forest classifier

| Hyperparameter | Amended value |
|---|---|
| C | 10 |
| penalty | l2 |
| solver | liblinear |

Table 14: The hyperparameters used in training the Logistic Regression classifier

| Hyperparameter | Amended value |
|---|---|
| C | 0.1 |
| kernel | linear |

Table 15: The hyperparameters used in training the Support Vector Machine classifier

## Appendix F Linguistic analysis

| Conversational AI Agent | Output |
|---|---|
| PaLM-2 | Workers choosing between warmth and a full refrigerator. Dreams put on hold, replaced by the daily grind of survival. The cost of living crisis is a call to action. It demands solutions, not platitudes. It requires bold leadership, targeted support, and a commitment to building a fairer, more resilient society. This isn't just a winter chill; it's a systemic squeeze. And until we collectively address its root causes, the UK's future risks being defined not by prosperity, but by the struggle to merely exist. |
| GPT-3.5 | The United Kingdom is currently contending with a profound cost of living crisis, as citizens grapple with the escalating prices of essential goods and services. From surging energy costs to rising food and housing expenses, the financial strain on households has become palpable. This crisis not only impacts individual wallets but also raises concerns about broader economic inequality. Many individuals and families are forced to reassess their budgets and make difficult choices to navigate through these challenging times. As the cost of living continues to rise, policymakers face the imperative of implementing effective strategies to alleviate the burden on citizens and foster economic resilience. The cost of living crisis in the UK is a pressing issue that demands thoughtful and comprehensive solutions to ensure the well-being of the population. |

Table 16: The GPT-3.5 (Text one (T1)) and PaLM-2 (Text two (T2)) output for the prompt *"Write me a write a short paragraph on the cost of living crisis in the UK"*

| Linguistic Features | Feature |
|---|---|
| Average word length | 5.125 charaters per word |
| Average sentence length | 14.9 words per sentence |
| Type-token ratio | 0.678 |
| Text repetitiveness rate | 0.32 |
| Character unigram | 'e' (occurs 98 times), 't' (occurs 78 times), 's' (occurs 58 times),<br>'i' (occurs 54 times), 'n' (occurs 54 times) |
| Character bigram | 'th' (occurs 45 times), 'es' (occurs 38 times), 'nt' (occurs 30 times),<br>'in' (occurs 28 times), 'er' (occurs 28 times) |
| Character trigram | 'the' (occurs 23 times), 'ing' (occurs 16 times), 'ion' (occurs 14 times),<br>'ent' (occurs 13 times), 'ndi' (occurs 9 times) |
| Sentence type | Simple (3), compound (2), complex (2), declarative (6),<br>passive (1), exclamatory (1) |
| POS and Dependency tags | Top 5 POS Tags:<br>NN (Noun, singular or mass), IN (Preposition or subordinating conjunction),<br>JJ (Adjective), VBZ (Verb, 3rd person singular present), DT (Determiner)<br><br>Top 5 Dependency Tags:<br>nsubj (Nominal subject), ROOT (Root of the clause), prep (Prepositional modifier),<br>pobj (Object of preposition), det (Determiner)<br><br>Number of shared POS tags: 5<br>Number of shared dependency tags: 5<br><br>Number of Unique POS tags: 13<br>Number of Unique dependency tags: 8 |

Table 17: Subset of features extracted from GPT-3.5 (T1) for linguistic analysis

| Linguistic Features | Feature count |
|---|---|
| Average word length | 5.276 characters per word |
| Average sentence length | 16 words per sentence |
| Type-token ratio | 0.607 |
| Text repetitiveness rate | 0.392 |
| Character unigram | 'e' (occurs 50 times), 't' (occurs 43 times), 's' (occurs 37 times),<br>'r' (occurs 30 times), 'i' (occurs 28 times) |
| Character bigram | 'th' (occurs 24 times), 'es' (occurs 23 times), 'ti' (occurs 21 times),<br>'in' (occurs 21 times), 're' (occurs 18 times) |
| Character trigram | 'the' (occurs 14 times), 'ion' (occurs 13 times), 'ing' (occurs 12 times)<br>'ent' (occurs 10 times), 'tio' (occurs 9 times) |
| Sentence type | Simple (5), compound (3), complex (4), declarative (12), passive (1) |
| POS and Dependency tags | Top 5 POS Tags:<br>NN (Noun, singular or mass), VBZ (Verb, 3rd person singular present),<br>IN (Preposition or subordinating conjunction), DT (Determiner),<br>JJ (Adjective)<br><br>Top 5 Dependency Tags:<br>nsubj (Nominal subject), ROOT (Root of the clause),<br>prep (Prepositional modifier), pobj (Object of preposition),<br>det (Determiner)<br><br>Number of shared POS tags: 5<br>Number of shared dependency tags: 5<br><br>Number of Unique POS tags: 21<br>Number of Unique dependency tags: 8 |

Table 18: Subset of features extracted from PaLM-2 (T2) for linguistic analysis