Probing the Uniquely Identifiable Linguistic Patterns of Conversational AI Agents

Anonymous ACL submission

Abstract

The proliferation of Conversational AI agents (CAAs) has emphasised the need to distinguish between human and machine-generated texts, with implications spanning digital forensics and cybersecurity. While prior research primarily focussed on distinguishing human from machine-generated text, our study takes a more refined approach by analysing different CAAs. We construct linguistic profiles for five CAAs, aiming to identify Uniquely Identifiable Linguistic Patterns (UILPs) for each model using authorship attribution techniques. Authorship attribution (AA) is the task of identifying the author of an unknown text from a pool of known authors (Juola, 2008). Our research seeks to answer crucial questions about the existence of UILPs in CAAs, the linguistic overlap between various text types generated by these models, and the feasibility of Authorship Attribution (AA) for CAAs based on UILPs. Promisingly, we are able to attribute CAAs based on their original texts with a weighted F1-score of 96.94%. Further, we are able to attribute CAAs 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), we identify the top 100 most informative linguistic features for each CAA, achieving a weighted F1-score ranging from 86.04% to 97.93%, and an overall weighted F1-score of 93.86%.

1 Introduction

Recent advances in deep learning and natural language processing have led to the emergence of conversational AI agents (CAA), which we define as large language models (LLMs) that can generate natural language as a dialogue system would. These have been applied in tasks such as question answering and text summarisation (Zhao et al., 2023). The widespread use of CAAs has highlighted the importance of determining the origin of a text (Desaire et al., 2023; Fagni et al., 2021; Mitrović et al., 2023). Authorship attribution for CAAs, i.e., the ability to ascertain the authorship of texts generated by CAAs, is crucially important in the area of user protection (e.g., the prevention of online hate crimes or distribution of misinformation) and academic malpractice (Mahmood et al., 2019). This arises due to the increasing popularity of CAAs (Desaire et al., 2023), which can be used as an obfuscation tool, allowing users to hide their writing style and spread potentially harmful content anonymously with the use of CAAs. This can be mitigated by building methods for CAA attribution: the task of identifying the CAA responsible for producing written text. Furthermore, it is important for such methods to reliably attribute texts to the corresponding CAAs that produced them, even if the texts were generated for different textual genres and thus follow different writing styles.

Prior research has predominantly focussed on distinguishing between human and machine-generated text (Fagni et al., 2021; Mitrović et al., 2023; Becker et al., 2023; Islam et al., 2023a; Markowitz et al., 2023), paying little attention to the investigation of different CAAs. 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 CAAs. As a result, we have meticulously crafted linguistic profiles 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 discern the presence of UILPs. We use these UILPs to perform authorship attribution (AA), which involves analysing features to identify patterns that can help distinguish between texts written by different authors (Juola, 2008, 2006; Sari, 2018). Analysing the discernible patterns in the writing of each CAA is crucial in enabling CAA attribution, regardless of the text it generates. We propose a transparent means for linguistic analysis that is more interpretable across different CAAs and forms the central emphasis of this paper.

This research area is novel and has yet to be explored. As aforementioned, there have been many attempts to identify texts generated by machines and humans, however, there has been no investigation on the UILP of CAAs, no comparison of different CAAs and no research indicating if these CAAs can be differentiated from each other based on their linguistic patterns. Moreover, there is a notable absence of analysis of CAAs based on *stylometry*, i.e., the statistical analysis of language often used in the context of forensic linguistics (Rocha et al., 2016). 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 CAAs based on their original texts, through the recognition of their UILPs?

RQ2: Can we attribute text to CAAs 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 CAA 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 CAAs, while the second dataset is an expanded version of the first whereby each text was paraphrased by the CAAs 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 CAA 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 CAAs based on principal component analysis (PCA).

2 Related Work

The field of AA encompasses three distinct categories, as outlined by Juola (2008). The first category pertains to closed-set attribution, where the objective is to identify the author of a text of an unknown text from a known pool of authors (Juola, 2006). The other categories are authorship verification and author profiling. In the case of verification case true author may not be in the list of suspected authors and the main challenge is to verify whether the suspected author is the author of a document or not. Profiling is the case of providing as much information about the author from a set of texts. Information such as their age, education level or gender, all of which can be seen in their use of linguistic devices (Sari, 2018). Our work is concerned with closed-set attribution.

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 the basis of AA (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). However, this theory has not been investigated in the case of CAAs, which is what we sought to achieve in our work.

Previous research on CAAs has primarily focussed on only the GPT family of 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), or surface-level linguistic features (Desaire et al., 2023; Markowitz et al., 2023). These studies lack a comparative analysis of various CAAs and do not incorporate any stylometric analysis in their evaluation, which would better capture the use of CAAs in generating texts in other scenarios. Other research demonstrates that human participants were unable to distinguish between texts written by humans and machines (Islam et al., 2023b; Cox, 2005).

[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/

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 CAAs 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 Methodology

Different CAAs may exhibit diverse approaches to conversation. By detecting these difference we allow used and developers to understand the specific characteristics of each CAA. This section details how the CAAs were selected, the data collection steps and our approach to CAA attribution.

3.1 Model Selection

The models used for this project include GPT-3.5, GPT-4, Text-Curie-001, PaLM-2 and, LlaMA2-7b. 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)[1] 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)[2] 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[3] (Language Learning and Meaning Acquisition) 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.

These models, all created by various developers, are widely used, with GPT being particularly prominent (Leiter et al., 2023). Our objective is to conduct a linguistic comparative study and to investigate whether these models, irrespective of their shared training methods, can exhibit unique patterns in their generated texts. Due to similarities in the manner in which they were trained, we can anticipate that these CAAs should, in theory, lack a significant difference in their UILPs, which could make them difficult to distinguish from each other.

3.2 Data collection

Our collection of CAA-generated texts 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[4] 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. The rationale for selecting these article topics was to ensure a diversity of texts within the dataset. For instance, within the education category, the most prominent headline pertained to the impact of Covid-19 anxieties on academic studies. Table 13 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 14 in B. These prompts were given as input to all the CAAs, which generated responses. Data collection occurred 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 (Touvron et al., 2023) and the GPT models (OpenAI, 2023). For each of the 10 prompts, 20 texts were generated. Thus, overall, 200 texts were generated per model except PaLM-2. The data for PaLM-2 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 second phase pertains to the collection of stylistic data for only GPT 3.5, 4 and Text-Curie-001 (OpenAI, 2023). We employed only these three CAAs because they responded effectively to the prompt, while other CAAs produced nonsensical texts or simply repeated text. The stylistic data uses the original data to produce paraphrases of this text in different stylistic genres. Firstly, we asked

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

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

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

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

each model to paraphrase the original text in a general manner, i.e., without specifying a specific style. The model is 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. 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 an 80:20 partition. No cleaning or preprocessing steps were applied to the data.[5]

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 major concern in this work as our aim is to focus on context-independent linguistic features.

3.3 Writeprints as Feature Representation

Abbasi and Chen (2008) proposed the Writeprint: a set of linguistic features for representing the distinctive writing style of each author of interest in an AA task. The said 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 CAAs 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 CAAs of interest (Section 3.4) and to train traditional machine learning-based classification models for attributing a given text to any of the CAAs (Section 3.5).

3.4 Analysing the UILP of CAAs

We employed principal component analysis (PCA) (Jolliffe and Cadima, 2016) to assess the top 100 most informative linguistic features that represent each model (based on its generated texts), as well as 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, and later on also investigated the top 200 and 300 features in a similar manner.

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.

3.5 Classification Models for AA

We cast AA 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 CAAs.

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 3.3, using both default parameter values and optimised parameter values. Optimised parameter values are defined through the use of GridSearchCV [6]. We use both default and optimised hyperparameters (optimised parameter values can be seen in Appendix **??**) to set a baseline and assess performance, enabling us to quantify the extent of improvements. The consistent superiority of optimised parameters indicates a robust and

[5] The datasets will be made publicly available upon paper acceptance

[6] GridSearchCV: `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

dependable model. To further strengthen this robustness, we compute the standard deviation (SD). Our results show that the SD in all experiments is low, indicating that the data points cluster closely around the mean. This consistency highlights the result's reliability.

Additionally, we sought to compare the AA performance of the above-mentioned traditional machine learning-based models with a transformer-based language model (TLM) (Vaswani et al., 2017), given that TLMs have shown superior performance in classification tasks including those in the area of digital forensics (Fabien et al., 2020). In this case, we selected the Decoding-enhanced BERT with Disentangled Attention (DeBERTa) model as it has been demonstrated to outperform other transformer models in a variety of tasks (He et al., 2021). We employed both a default hyperparameter DeBERTa model as well as a finetuned model. The DeBERTa model was fine-tuned for our task using our datasets and was trained over the course of 6 epochs; further details for the ML and TLM models can be found in Appendix D. 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 over a single run.

4 Evaluation Results and Discussion

In this section, we discuss how the results align with each research question and if the results support the existence of a UILP in each CAA.

4.1 Attribution of CAA Original Texts

Table 2 and table 3 present the results for the AA of the original data. The EWP features were extracted from all the texts and the methodology was applied, as described in Section 3.3. 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 5.78% demonstrating competitive performance. When the extended feature set is combined with an ML classifier, the weighted F1-scores ranged from 93.33% to 94.88% when default hyperparameters were used and 93.88% to 96.94% when the model was optimised. This demonstrates that each CAA does have a UILP as we can attribute each model to the correct CAA with a weighted F1-score of at least 93.33%.

ML Model	Accuracy	W-F1	SD
SVM [d]	93.56	93.33	0.19
RF [d]	96.14	95.02	0.69
LR [d]	94.86	94.88	0.04
SVM	93.87	93.88	0.00
RF	96.54	96.54	0.37
LR	96.94	96.94	0.00

Table 2: Performance Metrics for Original Data Attribution: the average Accuracy, Weighted F1-score (W-F1) and Standard deviation Scores for optimised and default [d] SVM, LR, RF classifiers (after 5 runs) for all CAAs

Model	Accuracy	W-F1
DeBERTa	99.11	99.11
DeBERTa [d]	98.43	98.41

Table 3: Performance Metrics for the original Data Attribution: the Accuracy and Weighted F1-score (W-F1) for a fine-tuned and default [d] DeBERTa model

From the results in Table 4 and Table 5, 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 experiment displays an impressively competitive performance with the optimised LR model having a weighted F1-score of 97.50%, only a 1.61% drop in the weighted F1-score when compared to a fine-tuned DeBERTa model.

4.2 Attribution of CAA Stylistic Texts

Apart from AA of the original data, we also investigated AA of stylistic text; this can be considered as

ML Model	Accuracy	W-F1	SD
SVM [d]	94.33	94.17	0.68
RF [d]	95.87	95.88	1.23
LR [d]	96.51	96.45	0.32
SVM	94.11	94.17	0.00
RF	96.67	96.67	0.00
LR	97.50	97.50	0.19

Table 4: Performance Metrics for the Attribution of all GPT datasets: the average Accuracy, Weighted F1-score (W-F1) and Standard deviation Scores for optimised and default [d] SVM, LR, RF classifiers (after 5 runs) for GPT-4, GPT-3.5 and, Text-Curie-001

Model	Accuracy	W-F1
DeBERTa	99.11	99.11
DeBERTa [d]	98.29	98.33

Table 5: Performance Metrics for the GPT Data Attribution: the Accuracy and Weighted F1 (W-F1) for a fine-tuned and default [d] DeBERTa model

ML Model	Accuracy	Weighted F1	SD
SVM [d]	75.28	75.43	0.43
RF [d]	78.28	77.94	0.26
LR [d]	75.14	75.26	0.10
SVM	95.56	95.56	0.00
RF	95.25	95.24	0.25
LR	95.83	95.84	0.00

Table 6: Performance Metrics for the Attribution of the Stylistic data: the average Accuracy, Weighted F1-score (W-F1) and Standard deviation Scores for optimised and default [d] SVM, LR, RF classifiers (after 5 runs) for GPT-4, GPT-3.5 and, Text-Curie-001

Model	Accuracy	Weighted F1
DeBERTa	88.00	88.00
DeBERTa-1	79.41	79.72

Table 7: Performance Metrics for the stylistic Data Attribution: the Accuracy and Weighted F1-score (W-F1) for a fine-tuned and default (-1) DeBERTa model

cross-genre attribution as we examine the attribution success of the same CAAs on different stylistic data.

The results of the AA of the stylistic dataset for GPT models are presented in Table 6 and Table 7. 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 CAA. The results here support the notion of the UILP existing in the different stylistic genres of 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, but 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. One can observe a 1.1% weighted F1-score drop when using an optimised LR model and a 19.62% drop when comparing the performance of default LR. These results indicate that each CAA has a distinct UILP for the stylistic texts, further affirming the idea that performance decreases across genres due to varying linguistic patterns (Stamatatos, 2016).

To conclude, we can recognise each CAA, regardless of the text's style, with the highest weighted F1-score achieved at 95.83%.

4.3 Principal Component Analysis of CAA

In this section, we identify the top 100 most informative linguistic features across all CAAs as well as the top 100 most informative linguistic features for each CAA; we then assess the extent to which attribution can be performed based on these features, for both original and stylistic data.

PCA is a statistical technique used for dimensionality reduction and is used to preserve the most important information. For all the original data, we extracted our Extended Writeprint 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. The outcomes of this analysis are presented in Tables 8 and 9.

When performing attribution using only the top 100 most informative linguistic features as extracted for all the original data (see Tables 8 and 9), we found that Text-Curie-001 has the highest weighted F1-score when using the top 100 features for any model and has a self-identifying weighted F1-score of 98.77% using an optimised LR model. LLaMA2-7b obtained the lowest weighted F1-score when being identified using its own top 100 features at 66.67%. The variation in the results in this table supports the idea of a UILP. When looking at the same 100 features for each CAA, the success in attributing the authors varies with a difference ranging from 66.67% to 98.77%.

These results support the theory of linguistic individuality (Nini, 2023) as the CAAs do not have identical grammars even though the training material, methods, the 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%. It is evident that each CAA struggles to distinguish itself when using its own top 100 most informative features. However, this is due to the substantial overlap in these features,

CAA	Accuracy	W-F1	SD
GPT-3.5	91.66	88.25	0.01
GPT-4	95.34	93.33	0.02
LLaMA2-7b	100	97.85	0.00
PaLM-2	89.13	87.23	0.01
Text-Curie-001	100	96.93	0.00
All	94.40	94.90	0.00

Table 8: Results of attribution using an LR model with default hyperparameters trained on the top 100 most informative linguistic features extracted using PCA across all datasets

CAA	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 9: Results of attribution using an optimised LR model trained on the top 100 most informative linguistic features extracted using PCA across all datasets

CAA	Accuracy	W-F1	SD
GPT-3.5	86.43	85.17	0.01
GPT-4	85.00	85.02	0.01
LLaMA2-7b	88.89	100	0.01
PaLM-2	91.14	83.72	0.00
Text-Curie-001	100	100	0.00
All	90.31	90.23	0.00

Table 10: Accuracy and weighted F1-score for each CAA when performing AA using only their unique features

CAA	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 11: Accuracy and weighted F1-score for each CAA when performing AA using only their unique features

as demonstrated in Appendix F. On average, they share more than 50% of their top 100 features with another CAA. This clarifies why, in Table 12, we observe an absence of a distinct pattern in CAAs' ability to identify themselves through their own 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, have similar training processes and serve the same purpose. GPT-4 is an improvement that builds upon the existing capabilities of GPT-3.5.

Further investigation was performed to delve into the subtle linguistic differences and to determine if CAAs can be identified based on their unique feature sets. We conducted a comparison of the top 100 features across all CAAs and identified features unique to each model. After obtaining the set of distinctive features for each model from this comparison, 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 X, Y, and Z as they were uniquely associated with GPT-4 in its top 100 most informative features. 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. The differences in results were significant: the weighted F1-scores ranged from 83.72% to 100% when using the default parameters of a model. This changed to 86.17%-100% when we optimised the hyperparameters (see Table 10 and 11). The results support the theory that when investigating a CAA's inherently unique features, one can attribute each CAA with greater success. Further results on the attribution success for each model can be seen in Tables 10 and 11.

The subsequent phase involved conducting PCA for each model and extracting the most informative top 100 features. Following this, we attempted authorship attribution for all models using these top 100 features, and the outcomes are presented in Table 12. The results indicate that only LLaMA2-7b could successfully self-identify as the most similar CAA 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,

CAA	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 12: Table displaying 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

the diagrams in Figure 1a in Appendix F illustrate substantial feature overlap among different models when analysing 300 features. However, as the features are reduced to find the most unique ones, there is a noticeable drop in overlap; see Figure 1b and Figure 1c in Appendix F. 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 RF algorithm.

5 Conclusion and future work

In our study, we have addressed three key research questions. Firstly, we have confirmed the presence of Uniquely Identifiable Linguistic Patterns (UILPs) in conversational AI agents (CAAs). This is supported by high accuracy in attribution success for both original and stylistic data, with weighted F1-scores ranging from 93.33% to 96.96% using features from our Extended Writeprint (EWP) feature set and traditional machine learning-based classifiers. We also demonstrate similar performance 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 AA performance compared to a fine-tuned DeBERTa model. Through PCA analysis, we explored the attribution of CAAs based on their UILPs and performed AA using these linguistic features. Our results show that the combination of our EWP and RF classification effectively supports cross-genre AA, with weighted F1-scores ranging from 94.17% to 97.50% for the AA of the stylistic data. This affirms the principle of linguistic individuality in CAAs, showcasing their UILPs. These findings validate the existence of UILPs in CAAs and offer valuable insights into their distinctive linguistic patterns, with potential applications in digital forensics, detecting fake news and plagiarism.

Future work will look to 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 CAAs both in English and cross-lingually.

6 Limitations

In our study, text generation using various APIs that make our CAAs 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 CAA outputs did not produce cohesive texts (in the case of LLaMA2-7b) or, produced very short texts (in the case of Text-Curie-001). Further, only a set of three text genres were investigated: academic articles, fictitious narratives, and tweets and social media posts (the latter most two falling under the same genre). To perform cross-genre AA we must expand this scope to cover a wider array of genres as well as investigate at different levels of formality.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems*, 26(2).

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.

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.

Michael T. Cox. 2005. Metacognition in computation: A selected research review. *Artificial Intelligence*, 169(2):104–141. Special Review Issue.

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.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023a. Distinguishing human generated text from chatgpt generated text using machine learning.

Niful Islam, Debopom Sutradhar, Humaira Noor, Jarin Tasnim Raya, Monowara Tabassum Maisha, and Dewan Md Farid. 2023b. Distinguishing human generated text from chatgpt generated text using machine learning.

Ian T. Jolliffe and Jorge Cadima. 2016. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374.

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

Patrick Juola. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval*, 1:233–334.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2023. Chatgpt: A meta-analysis after 2.5 months.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*, 2019:54 – 71.

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*.

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.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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 CAAs

	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 13: The prompts used to collect CAA-generated texts. All original texts were produced all by CAAs: 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 14: 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
		Sentence 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 15: 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 16: 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 17: The hyperparameters used in training the Random Forest classifier

Hyperparameter	Amended value
`C`	10
`penalty`	l2
`solver`	liblinear

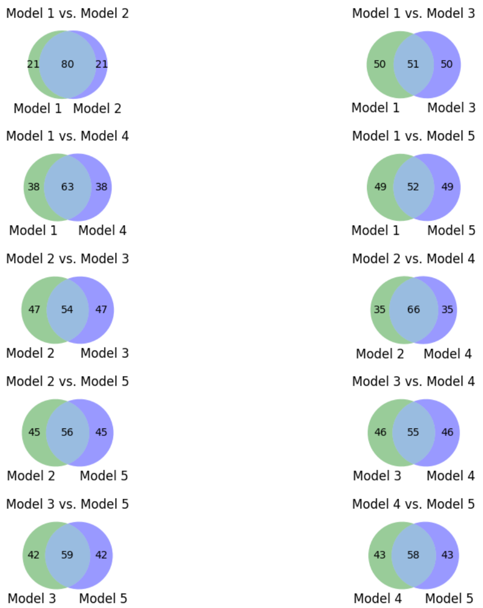
Table 18: The hyperparameters used in training the Logistic Regression classifier

Hyperparameter	Amended value
`C`	0.1
`kernel`	linear

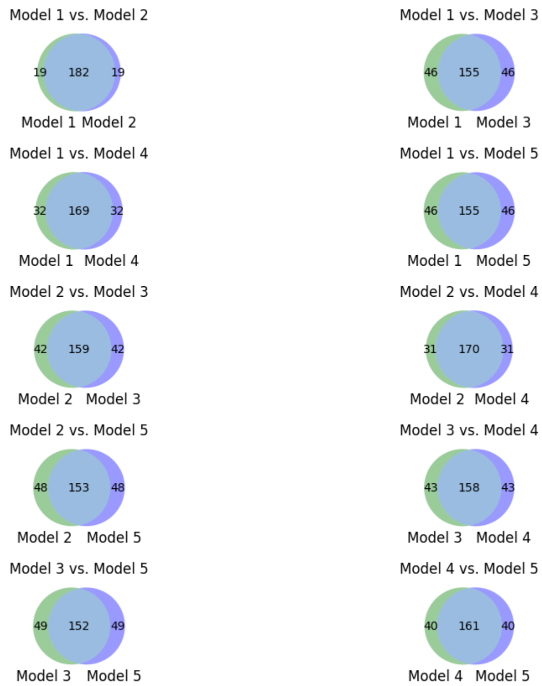
Table 19: The hyperparameters used in training the Support Vector Machine classifier

Appendix F PCA visualisations

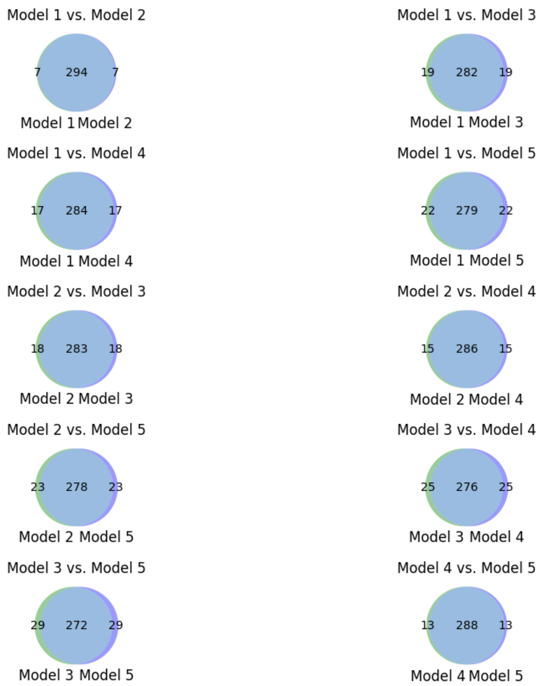
Key:
Model 1: GPT-3.5; Model 2: GPT-4; Model 3: LLaMA2-7b; Model 4: PaLM-2; Model 5: Text-Curie-001.



(a) Overlap for the top 100 features of all CAAs



(b) Overlap for the top 200 features of all CAAs



(c) Overlap for the top 300 features of all CAAs

Figure 1: Overlap for the top 200 most informative linguistic features extracted based on our EWP using PCA for all CAAs. Classification results are in Table 12