scientific reports

OPEN

Examining embedded lies through computational text analysis

Riccardo Loconte[1,2] & Bennett Kleinberg[2,3]

Verbal deception detection research relies on narratives and commonly assumes statements as truthful or deceptive. A more realistic perspective acknowledges that the veracity of statements exists on a continuum, with truthful and deceptive parts being embedded within the same statement. However, research on embedded lies has been lagging behind. We collected a novel dataset of 2,088 truthful and deceptive statements with annotated embedded lies. Using a counterbalanced within-subjects design, participants provided two versions of an autobiographical event. One was described truthfully, and the other one deceptively by including embedded lies. Participants later highlighted those embedded lies and judged them on lie centrality, deceptiveness, and source. We show that a fine-tuned language model (Llama-3-8B) can classify truthful statements and those containing embedded lies significantly above the chance level (64% accuracy). Individual differences, linguistic properties, and explainability analysis suggest that the challenge of moving the dial towards embedded lies stems from their resemblance to truthful statements. Typical deceptive statements consisted of 2/3 truthful information and 1/3 embedded lies, largely derived from past personal experiences and with minimal linguistic differences from their truthful counterparts. We present this dataset as a novel resource to address this challenge and foster research on embedded lies in verbal deception detection.

Keywords Deception, Embedded lies, Lying profile, Natural Language processing, Individual differences

Everyone engages in some form of deception daily[1]. Rather than fabricating entirely false accounts, however, most individuals tend to combine elements of truth with elements of falsehood[2]. This deception strategy is known as the embedding of lies. Embedded lies present a distinctive challenge in deception research and remain a largely under-investigated phenomenon.

Verbal deception detection

Research on verbal deception detection has often been focused on methods using manual coding, which involve training human judges to evaluate statements based on predefined verbal cues. One of the most common and widely applied techniques in real-world settings is the Criteria-Based Content Analysis (CBCA)[3]. CBCA was originally developed to evaluate children's testimonies on alleged sexual abuse cases and is now used to assess the credibility of testimonies in legal contexts. CBCA requires a human to identify and score a narrative on specific verbal cues, such as the amount of detail, unexpected complications, or spontaneous corrections, that truth-tellers are more likely to exhibit than deceivers. Another widely investigated technique in research is Reality Monitoring (RM)[4,5], which distinguishes between truth and lies by focusing on the richness of sensory and contextual details provided by the speaker. Truth-tellers are thought to provide more vivid and detailed sensory information than deceivers, who typically rely on fabricated or imagined events. Building on the RM, the Verifiability Approach (VA)[6] capitalises on the tendency of truth-tellers to provide more verifiable details compared to lie-tellers, who avoid that because it could expose their deceit. While these methods have promise[3,7–9], they are more time-consuming and reliant on the expertise of practitioners than automated procedures with computational models, limiting their scalability[10].

Computer-automated verbal deception detection

Recent advances in Natural Language Processing (NLP), often combined with methods from machine learning (ML), have introduced automated methods for detecting deception, enhancing both scalability and objectivity. NLP techniques allow the representation of textual data in a numerical vector form, with different levels of granularity. For instance, the Linguistic Inquiry and Word Count (LIWC)[11] computes the frequency of words that pertain to psychological, social, and emotional dimensions (e.g., cognitive words, affective words, social words, etc.); part-of-speech (POS) tagging informs on the shallow syntactic text structure by automatically

[1]Molecular Mind Lab, IMT School of Advanced Studies Lucca, Lucca, Italy. [2]Department of Methodology and Statistics, Tilburg University, Tilburg, Netherlands. [3]Department of Security and Crime Science, University College London, London, UK. email: riccardo.loconte@imtlucca.it

assigning grammatical categories (e.g., nouns, verbs, pronouns) to words; named-entity recognition (NER) identifies and labels proper nouns and piece of information into predefined broader categories (i.e., *25.12.2024* into DATES, *Central Park* into LOCATIONS, *Google* into ORGANIZATIONS); *n*-gram models represents the text into frequency patterns of tokens; and embeddings use a vectorial numerical representation of words or statements that preserves the semantic and contextual relationship, allowing similar items to have closer representations in a multi-dimensional space.

Early approaches in computer-automated verbal deception detection (e.g[12–15])., first extract information from textual data and then use supervised machine learning to train models that use these extracted variables to derive a truthfulness judgment (see[10,16,17] for an extensive overview on the topic). For example, a previous study detected deceptive opinions by training a naïve Bayes and a support vector machine classifier on n-grams reaching 70.8% and 70.1% accuracy[12]. Likewise, another study detected deceptive opinions after training a support vector machine classifier on a combination of n-grams and LIWC features, reaching 89.8% accuracy[13]. Finally, other scholars extracted the proportion of unique named-entities and found a significant discriminative power of 0.67 and 0.65 in detecting positive and negative deceptive hotel reviews, respectively[14].

In contrast, more recent approaches go beyond feature extraction and use recent advances with Transformer-based models[18] that already incorporate a semantic representation of texts. For example, a previous study employed a Bidirectional Encoder Representations for Transformers (BERT)[19] to incorporate contextual embeddings with attention-based mechanisms and detected deceptive utterances in a dataset of transcripts of criminal proceedings hearings, reaching 71.61% accuracy (DeCour)[20]. Another study fine-tuned a large language model (i.e., FLAN-T5)[22] to classify deceptive statements in three datasets encompassing personal opinions, autobiographical events, and future intentions, reaching 79.31% accuracy[22].

This increasing sophistication of NLP techniques paves the way for future research to further refine methods and address more complex challenges in automated deception detection, such as the detection of embedded lies.

Embedded deception

The concept of embedded lies emerged when researchers started to ask lie-tellers about their lying strategies. Across a range of different contexts and studies, lie-tellers were found to frequently draw on past experiences to make their lies more believable[23–27]. For instance, in two studies, 67% and 86% of liars, respectively, chose to construct their deceptive statements by incorporating elements of previously experienced events[28]. Similarly, in another research, over 20% of deceptive statements consisted of truthful information[6].

Despite being widely acknowledged within the field of deception detection, only a limited number of studies have directed their attention towards the phenomenon of embedded lies. Most research has focused on fabrication, conceptualising fabricated stories as entirely false and resulting in a dichotomous view of deception as either completely deceptive or not deceptive at all. A study typical of this perspective research requires a between-subjects design where participants engage in or view a mock crime event and are then allocated to one of two conditions: truth-tellers, in which they recollect exactly what they watched or experienced, and lie-tellers, in which they fabricate the event. For example, in a previous study, truth-tellers played a game during a staged event with a confederate, while lie-tellers -who did not participate in the event- were instructed to steal £10 from a wallet and then fabricated their involvement in the staged activities during a subsequent interview[29]. In other experiments, researchers implemented a matched-pairs design to match participants with the specific content of false statements[30–32]. For example, participants in the honest condition were asked to tell the researcher about their past holidays, whereas, in the deceptive condition, participants were instructed to pretend to have experienced that same holiday[32].

However, a more nuanced and realistic perspective acknowledges that a deceptive statement exists on a continuum where truthful and deceptive parts are embedded within the same statement (see Fig. 1). We, hence, adopted this framework in this paper to address the challenge of moving the dial towards embedded lies.

Detecting embedded lies

While embedded deception has been acknowledged as a problem by many[6,23–28], only a few studies reported on the nature and detection of embedded lies.

Using manual coding methods, two studies investigated embedded lies as fabricated statements contained within otherwise truthful statements[33,34]. The first found that lies embedded in a fabricated statement were not qualitatively different from lies embedded in an otherwise truthful statement, suggesting that verbal credibility assessment tools may be robust against the embedding of lies. The second was a follow-up study employing the within-statement baseline comparison[35], which is a strategy based on the idea that deception can be better detected if compared with a truthful baseline. However, the results showed that this strategy was not effective in improving the deception detection rate.

Another work investigated embedded lies to test whether specific manually annotated verbal cues, such as complications, common knowledge details, and self-handicapping strategies, varied across different amounts of lying in statements[36]. The study found only a significant difference in the number of complications, with a larger difference between truthful and outright deceptive statements compared to truthful and embedded lies.

Another work challenged the dichotomy between bald-faced lies and bald-faced truths[2]. It was hypothesized that both truthful and deceptive narratives draw from a common pool of information so that lies and truths are not mutually exclusive but rather coexist within the same accounts. In this study, participants were randomly allocated to the truthful or deceptive condition and were tasked with writing opinions about their friends. They were then instructed to indicate within their statements the specific parts that were deceptive and those that were truthful. The results suggested that truthful narratives inherently include a certain proportion of embedded lies (37.06% of embedded lies in deceptive texts compared to 29.21% in truthful texts). While deceptive narratives

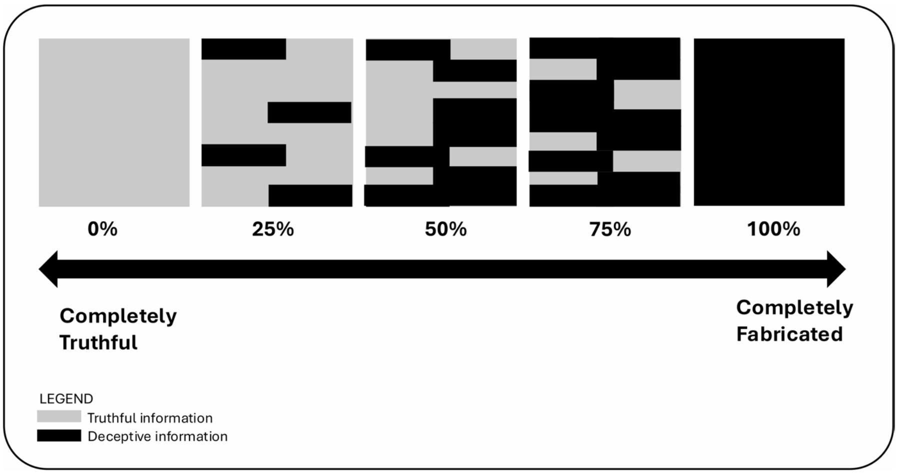

Fig. 1. Graphical representation of the deception continuum framework. Deception is embedded into truthful statements in a continuous range from 0% (= no lies are present) to 100% (= the whole statement is made up). Levels in between represent various degrees of deception in the form of embedded lies.

contained a significantly higher rate of fabrications than truthful narratives, it is noteworthy that even the latter was aplenty with fabrications.

The few studies to date have not yet addressed two important challenges in lie detection research. Firstly, measuring embedded lies is inherently complex, with few studies employing within-subjects designs that control for individual differences and statement topics. Secondly, the lack of granular analytical methods hampered the detection of embedded lies, which are harder to identify than general deception. These methodological limitations have hindered the exploration of embedded lies, leaving them underrepresented in the literature despite their significance. Additionally, previous scholars have mentioned the importance of individual differences (e.g., demographic factors, personality traits, cognitive styles, and emotional states) in engaging in deceptive behaviour and in the type and dynamic of the deception involved[37–44] (for a recent and complete review, see[45]). In the context of embedded lies, only one study explored whether and how personality (i.e., dark triad traits)[46] and demographic factors (i.e., age, gender, ethnicity, and political ideology) influence this specific form of deception[2]. However, no significant differences emerged from this specific study. Hence, with respect to individual differences, embedded lies represent an even more unexplored phenomenon. We, therefore, aim to connect these two streams of research in this paper by also promoting the investigation of individual differences in embedded lies.

The current study

This paper aims to help bridge the gap between deception practice and research by focusing explicitly on embedded lies. Prior work has usually employed between-subject or matched-pairs designs to study deception intended as fully fabricated accounts. Further, the majority of deception work relies on relatively small datasets[47] and manual procedures (e.g[48]).,. Embedded lies also need further investigation in terms of individual differences, with only one study focusing on demographic and individual traits affecting embedded deception[2]. We seek to address these limitations. First, we present a new dataset of embedded lies collected in a within-subjects experimental design that is sufficiently large to conduct meaningful computational analysis, including predictive modelling. Second, we enrich the scope of the dataset beyond the narratives and provide data at the individual level, allowing for analyses of individual differences in verbal deception behaviour. Third, we utilize automated approaches to retrieve variables from the narrative data using NLP methods and further resort to supervised machine learning to train models in detecting embedded deception.

Materials and methods

Ethics declarations

The study was approved by the Ethics Review Board of the Tilburg School of Social and Behavioral Sciences (Reference Number: TSB_RP1442). Data collection reported in this paper was conducted in accordance with the Declaration of Helsinki. Informed consent was obtained from all participants prior to their involvement in the study, and they were subsequently debriefed once their participation had been completed.

Participants

The sample size was determined by conducting an a priori power analysis for a small effect with a power of 0.90 (Cohen's $d = 0.20$, $\alpha = 0.05$, two-tailed), which resulted in a sample size of 265 participants. Since we aimed to present a dataset adequate for computational analyses, we collected significantly more data. We recruited a total of 1058 participants fluent in English from the general population through the online participant pool Prolific. Each participant provided informed consent before taking part in the Qualtrics-administered experimental task. Participation in the study was reimbursed 2$ upon experiment completion. Eight participants who did not follow the instructions (i.e., repeated the same phrase in multiple boxes) or provided non-sensical completions to the open-answer fields (e.g., writing random characters to fill the box) were removed for analysis. Eight participants from the subset of participants that freely recall a memorable event (after selecting the option "None of them") were removed because provided a title story that was too long (i.e., with a number of words higher than two standard deviations from the average) and were basically anticipating the main story in the wrong section. The final sample consisted of 1042 participants (58.23% females, 41.17% males, 0.19% preferred not to say, 0.38% expired data or removed consent on Prolific). The mean age was 30.32 years (SD = 9.35, range: 18–105).

Experimental task

A previous study found that truthful statements may contain deceptive parts[2]. However, we argue that truthful statements may also be, by definition, completely truthful, and those that are partially deceptive might be residuals from research design.

For this study, we developed an experimental task that followed a different perspective (see Fig. 1), considering fabrication on a continuum from 0 (fully truthful statements) to 100 (fully deceptive statements).

The experiment was conducted in a counterbalanced within-subjects design (Fig. 2) where half of the participants performed the truth-telling task and then the deceptive task, followed by the embedded lie selection and rating. The other half performed the deceptive task first, followed by the embedded lie selection and rating, and then the truth-telling task.

Additionally, we collected demographic variables and participants' lying attitudes to advance the understanding of individual differences in embedded lies.

Step 1: Event selection

The experiment started with the event selection task. Participants were provided with a list of eleven pre-selected autobiographical events that they might have experienced in the past 24 months. The events were deemed relevant for lying in the subsequent deceptive writing task. After participants selected all of the events that they had experienced themselves in the past 24 months, they were randomly assigned to one of them and answered five questions about the event with the aim of collecting the following memory-related variables:

(i) **time**: "*how long ago did the event happen?*" through a multiple-choice question with 25 options (from < 1 months to 24 months);
(ii) **recollection**: "*how often do you think or talk about this event?*" on a 5-point scale (1 = never; 5 = always);
(iii) **importance**: "*how important is this event to you?*" on a 5-point scale (1 = not important at all; 5 = extremely important);
(iv) **accuracy**: "*how well do you remember this event?*" on a 5-point scale (1 = not well at all; 5 = extremely well);
(v) **valence**: "*how would you rate this event in emotional terms?*" on a 5-point scale (−1 = extremely negative; −0.5 = somewhat negative; 0 = neither positive nor negative; 0.5 = somewhat positive; 1 = extremely positive).

The assigned event served as the basis for the remainder of the task. If participants did not experience any of the events in the list, they were instructed to choose the option "*none of them*". They were then asked to think about a positive or negative event occurring in the last 24 months that was memorable, emotional, and that directly

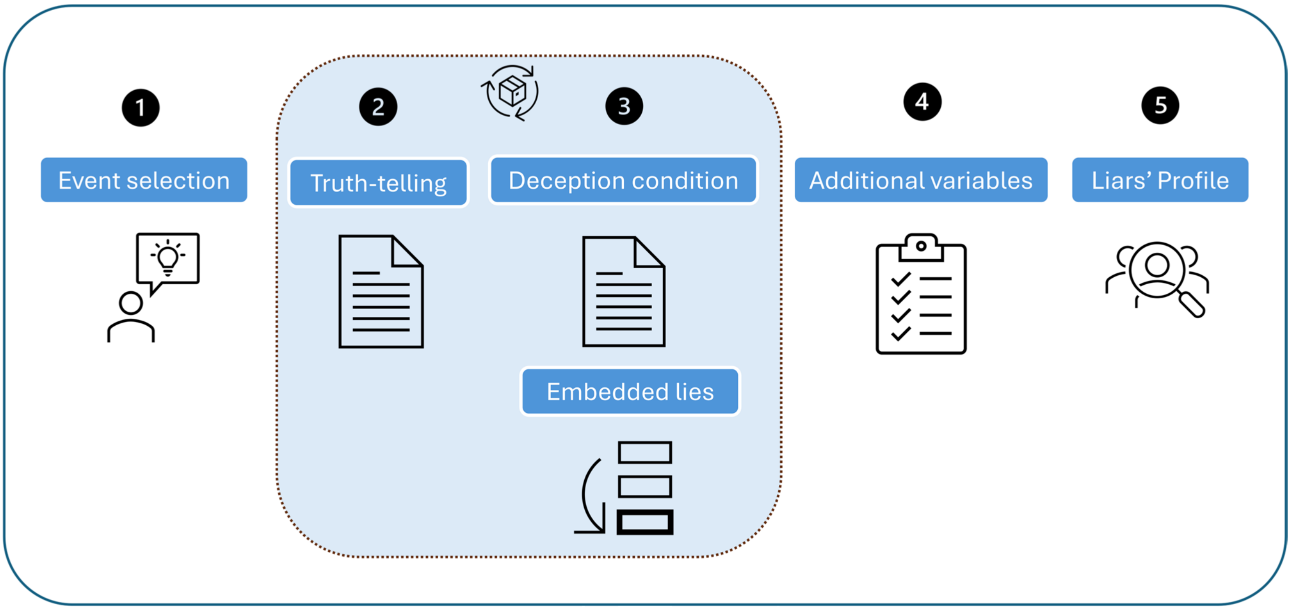

Fig. 2. Experimental procedure adopted in the experiment.

involved them and were asked to provide a short title for the event. The 18.91% ($n = 197$) of participants chose the "*none of them*" option.

We focused on autobiographical events that were deemed relevant for lying to mirror real-world scenarios and enhance the practical application of our research findings in improving the accuracy of deception detection.

Step 2: Truth-telling task

In the truth-telling task, participants were required to write an accurate and truthful account of the event in question. They were also asked to use correct spelling and grammar and were reminded not to use AI assistants in the writing task. Copy-pasting was blocked to prevent it. This task required a minimum of 3000 characters to move to the next phase of the experiment. Full instructions are provided in Supplementary Materials (SM) 1.

Step 3a: Deception task

For the deception task, participants were provided with a context relevant to lying and instructed to write a deceptive account of the selected event by incorporating false information. Specifically, participants were told to write an alternative version of the selected event in order to get a specific advantage from lying. Participants were also warned not to make up a statement about a new event and not to mention in any way that they were lying.

The list of contexts, matched for each event, and the number of participants allocated are provided in Table 1. Other than the deceptive instructions, the general writing instructions were identical to the truthful task. To motivate participants to do their best, they were informed about the chance of winning an extra 50£ if their statement was considered credible by the experts. In reality, all participants were included in the draw and the payment was distributed to a randomly selected participant after the data collection concluded. Full instructions are provided in SM 1.

Step 3b: Embedded lies

Once participants had written the deceptive account, participants were instructed to copy and paste words, phrases, or sentences from their deceptive statements into a maximum of 20 text boxes (similar to[2]). For each word or phrase that was copy-pasted, participants rated the deceptiveness (*"how deceptive was this detail for your whole lie statement?")* and centrality *("how central was this detail for your whole lie statement?"*) of each embedded lie on a 5-point scale (1 = not at all deceptive/central to 5 = extremely deceptive/central).

Through a multiple-choice question, participants provided the source on which they relied for the embedded lie. The source options were based on liars' relying on their past experiences or cognitive processes (i.e., from memory, imagination, and planning). The following source options were provided: (1) you connected the detail to a past personal experience; (2) you saw a similar event happen to someone else and used that as a basis for the detail; (3) you derived the detail from a story another person told you, or from a book, or a movie; (4) you imagined the detail without any specific memory or experience; (5) you used planned, future activities as a reference.

To account for individual variability (i.e., participants copy-pasting a single word vs. multiple phrases or sentences), the number of embedded lies was also standardized for each subject by computing the ratio between the number of words provided in the 20 boxes and the total number of words in their deceptive text. The standardized number of embedded lies ranged from 0 to 1.

Step 4: Additional variables

Once the two writing tasks were completed, participants rated the following additional variables on a 5-point scale (1 = completely disagree; 5 = completely agree): (i) difficulty of the task (i.e., "*I found the task was difficult*"); (ii) clarity of instructions (i.e., "*I found the instructions were clear*"); (iii) motivation of telling the truth (i.e., "*I was motivated to provide a convincing truthful statement*"); (iv) motivation of lying (i.e., "*I was motivated to provide a convincing deceptive statement*").

Events	Context for lying	No. of participants allocated (%)
A job interview for your dream job	Inflate your past experiences to get the job	160 (15.36%)
Being hospitalized and undergoing surgery	Exaggerate some side effects to receive extra compensation from your health insurance	70 (6.67%)
Being involved in a car accident	Increase the claimed amount of damage you received to get more money	47 (4.51%)
Causing a car accident	Describe the event so that it's not your fault	15 (1.44%)
Cheating on an exam	Describing how you passed the exam, given that you cannot admit that you cheated	48 (4.61%)
Cheating on your partner	Convince your partner that you didn't cheat on them	36 (3.45%)
Ending a long romantic relationship	Pretend that you just had an argument with your partner	152 (14.59%)
Getting a speeding fine	Pretend it wasn't you driving the car that day	62 (5.95%)
Getting fired	Pretend that you just had a bad day at work	34 (3.26%)
Missing a deadline at work because of bad organization	Find excuses that allow you not to appear forgetful or disorganised	97 (9.31%)
None of them	-	197 (18.91%)
Taking the bus/train without the ticket	Convince the ticket inspector that they shouldn't fine you for not having the ticket	124 (11.90%)

Table 1. List of events, contexts for lying, and number (percentages) of participants allocated to that event.

Step 5: Liars' profile

To measure potential individual differences in participants' lying attitudes, the lying profile questionnaire[49] was administered. The lying profile questionnaire measured dispositional traits of deception and was composed of 16 items grouped into four factors: frequency of lying (frequency); ability to lie (ability); negative attitude towards lying (negativity); and positive attitudes toward lying depending on the context (contextuality). Since participants may be prone to mask their lying attitude, the Balanced Inventory of Socially Desirable Responding Short Form (BIDR)[50] was also administered and used to correct the lying profile scores for potential effects of social desirability. The BIDR was a 16-item questionnaire which measured two main dimensions of social desirability: (1) self-deception enhancement (SDE): the unconscious tendency of individuals to provide honest but positively biased self-reports to protect self-esteem; (2) impression management (IM): the habitual and conscious tendency of individuals to present themselves of a favourable public image. We report results on both the raw lying profile scores as well as the ones after correcting for the BIDR scores. The correction procedure was conducted by fitting a general linear model that regressed out the SDE and IM scores from each lying profile factor.

Textual analysis of narrative data

Linguistic inquiry and word count analysis

The Linguistic Inquiry and Word Count (LIWC)[11,51] software is the gold standard for analysing word usage and semantics in texts across more than 100 features by calculating the percentage of total words corresponding to each category using validated dictionaries of words associated with psychosocial dimensions. Specifically, the English dictionary (version LIWC-22) was employed for this analysis, and 118 features were extracted from tokenized text.

DeCLaRatiVE stylometry

The DeCLaRatiVE stylometry approach[22] subsumes 26 linguistic variables derived from four theoretical lines in verbal deception research: Distancing[52], Cognitive Load[53,54], Reality Monitoring[4,55], and VErifiability Approach[7,56]. Linguistic variables associated with the cognitive load, such as text length, readability, and complexity, were computed using the Python library TEXTSTAT. Those related to the Distancing and RM framework were computed using LIWC-22 features[11,51] extracted from tokenized text. RM was also investigated through linguistic concreteness by cross-referencing an annotated dataset[57] with the content words in our dataset and averaging the final scores per statement. The preprocessing steps to derive content words from statements were tokenization, conversion to lowercase, stop-word removal, and lemmatization and were run with the SpaCy library in Python. Finally, verifiable details were extracted as entities with the named-entity recognition (NER) model available on the SpaCy library (en_core_web_trf, https://spacy.io/models/en#en_core_web_trf). A full list of the 26 linguistic variables with a short description is shown in Table 2 (refer to the original work[22] for a deeper understanding of the approach).

n-gram differentiation

Using the *n*-gram differentiation test[58], we compared the frequencies of unigrams, bigrams, and trigrams in truthful and deceptive statements within each event. This comparison was made using a signed rank sum test approach. Ties in ranks were fixed by averaging random ranks in 500 iterations. Statements were first preprocessed using SpaCy library in Python by removing stop words and lemmatising the remaining words. Only *n*-grams that appeared in at least 5% of all documents were included in the analysis. The effect size used for the frequency comparisons was *r*, which ranged from − 1.0 to 1.0.

Machine-learning classification

To investigate whether deceptive statements with embedded lies can be distinguished from truthful statements, we performed a document classification task using different state-of-the-art ML approaches. The models included both traditional and advanced architectures. Specifically, four Random Forest (RF) classifiers were trained on Bag of Words (BOW) representations[59], LIWC variables[11], DeCLaRatiVE variables[22], and GPT-embedding representations (https://platform.openai.com/docs/guides/embeddings), respectively. Additionally, we tested the performance of different fine-tuned language models, such as distilBERT[60], FLAN-T5 base[21], and Llama-3-8B[61]. We finally explored the performance of a deception language model from a previous study[22], which was a FLAN-T5 base model fine-tuned on three large datasets of deception with 79.31% (± 1.3) accuracy. Language models (i.e., distilBERT, FLAN-T5, and Llama-3) were trained using the HuggingFace library and the Google Colaboratory Pro + interface with the A100 Tensor Core GPU. Cross-validation was performed to ensure robust evaluation. Specifically, RF models were trained using 10-fold nested cross-validation, while language models were fine-tuned with 5-fold cross-validation to optimize computational costs. Classification performance was assessed in terms of overall accuracy (Formula 1), as well as precision, recall, and F1-score by condition (truthful vs. deceptive). Specifically, for each condition, precision measured the proportion of positive predictions that were actually positive (Formula 2); recall measured the proportion of actual positives that were correctly classified as positives (Formula 3); and F1 was the harmonic mean of precision and recall (Formula 4). Details of each model and the training procedure are reported in SM 4.

$$Accuracy = \frac{T_P + T_N}{T_P + F_P + T_N + F_N} \tag{1}$$

$$Precision = \frac{T_P}{T_P + F_P} \tag{2}$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (3)$$

$$F1\,score = \frac{2\ \cdot\ Precision\ \cdot\ Recall}{Precision\ +\ Recall} \quad (4)$$

where T_P = True positives, T_N = True negatives, F_P = False positives, and F_N = False negatives.

Analysis plan

We first looked at the subject level to examine characteristics of the reported embedded lies, such as their frequency, source, deceptiveness, and centrality for a deceptive account. Second, we examined individual differences related to demographic variables and lying profiles. Furthermore, we assessed linguistic differences between the narratives by using the LIWC and DeCLaRatiVE approach[22]. From the LIWC, we obtained for each subject 118 variables; from the DeCLaRatiVE stylometry technique, we obtained 26 variables. A within-subject permutation t-test with 9,999 permutations[62] was employed to test for statistical differences in these variables by statement veracity (truthful vs. deceptive). Results from multiple comparisons were corrected using Bonferroni correction. Truthful and deceptive statements were also analysed in terms of *n*-grams by using the *n*-grams differentiation test. These analyses were conducted in R using the *MKinfer* and *effectsize* libraries. Finally, state-of-the-art machine learning approaches were employed in a classification task to differentiate truthful from deceptive statements with embedded lies.

Results

Corpus descriptives

We collected a corpus of 2084 truthful and deceptive statements, collectively, across 11 events deemed relevant for lying. Descriptive statistics of the variables associated with the events are reported in Table 3S in SM 2. We found deceptive statements (M = 7.13, SD = 4.65) containing a significantly higher average number of sentences than truthful statements (*M* = 6.77, *SD* = 4.48), $t_{(9999)}$ = −0.37, *p* = .003, *d* = −0.08 [−0.14, −0.03]. Likewise, the number of words was, on average, significantly higher in deceptive (*M* = 145.29, *SD* = 83.65) than in truthful statements (*M* = 131.74, *SD* = 78.49), $t_{(9999)}$ = −13.55, *p* <.001, *d* = −0.17 [−0.22, −0.12]. However, these findings might be an artefact of the instructions, as participants were instructed to add details to appear deceptive and achieve a specific goal, resulting in producing longer statements.

Label	Description
num_sentences	Total number of sentences
num_words	Total number of words
num_syllables	Total number of syllables
avg_syllabes_per_word	Average number of syllables per word
fk_grade	Index of the grade level required to understand the text
fk_read	Index of the readability of the text
Analytic	LIWC summary statistic analyzing the style of the text in term of analytical thinking (0–100)
Authentic	LIWC summary statistic analyzing the style of the text in term of authenticity (0–100)
Tone	Standardized difference (0-100) of ‘tone_pos’ - ‘tone_neg’
tone_pos	Percentage of words related to a positive sentiment (LIWC dictionary)
tone_neg	Percentage of words related to a negative sentiment (LIWC dictionary)
Cognition	Percentage of words related to semantic domains of cognitive processes (LIWC dictionary)
memory	Percentage of words related to semantic domains of memory/forgetting (LIWC dictionary)
focuspast	Percentage of verbs and adverbs related to the past (LIWC dictionary)
focuspresent	Percentage of verbs and adverbs related to the present (LIWC dictionary)
focusfuture	Percentage of verbs and adverbs related to the future (LIWC dictionary)
Self-reference	Sum of LIWC categories ‘i’ + ‘we’
Other-reference	Sum of LIWC categories ‘shehe’ + ‘they’ + ‘you’
Perceptual details	Sum of LIWC categories ‘attention’ + ‘visual’ + ‘auditory’+ ‘feeling’
Contextual Embedding	Sum of LIWC categories ‘space’ + ‘motion’ + ‘time’
Reality Monitoring	Sum of Perceptual details + Contextual Embedding + Affect - Cognition
Concreteness score	Mean of concreteness score of words
People	Unique named-entities related to people: e.g., ‘Mary’, ‘Paul’, ‘Adam’
Temporal details	Unique named-entities related to time: e.g., ‘Monday’, ‘2:30 PM’, ‘Christmas’
Spatial details	Unique named-entities related to space: e.g., ‘airport’, ‘Tokyo’, ‘Central park’
Quantity details	Unique named-entities related to quantities: e.g., ‘20%’, ‘5 $’, ‘first’, ‘ten’, ‘100 meters’

Table 2. List and short description of the 26 linguistic features pertaining to the DeCLaRatiVE stylometry technique.

Embedded lies

Embedded lies included an average of 5.03 lies per text with an average number of words of 46.27 (*SD* = 42, *Median* = 35, see Table 3). The average ratio between the number of words in the annotated embedded lies and the respective deceptive statement was 0.32 (*SD* = 0.20, Median = 0.29). The average of embedded lies for each event is reported in Table 4S (SM 2). Using a 5-point scale, embedded lies were rated as moderately deceptive (*M* = 3.94, *SD* = 0.79, *Median* = 4) and central to the overall narrative (*M* = 3.55, *SD* = 0.82, *Median* = 3.59). Further, 35.86% of embedded lies (*n* = 1881) relied on personal past experiences that involved participants directly and 10.41% (*n* = 546) indirectly; 33.86% of embedded lies (*n* = 1776) relied on participants' imagination, while 14.95% (*n* = 784) on others' experiences and only 4.92% (*n* = 258) on personal future plans. An example of subjects' responses is provided in Box 1. Correlational analysis between variables associated with embedded lies, lying profile and BIDR scales is provided in SM 2.

EVENT: Being involved in a car accident	**EVENT: Being involved in a car accident** **INSTRUCTIONS: lie about the event to increase the claimed amount of damage you received to get more money**
«I was driving home after getting my dog from her sitter. My mom was sitting next to me to keep company to the dog, when we got met with a lot of traffic. So we were advancing quite slowly towards our destination when we come across this intersection, where on the right the cars have a STOP sign. This guy, very old, probably in his 60 s, doesn't stop and continues moving towards us. I stomped on the break, but it wasn't in time, and the car crashed against our side. It had been years since I was involved in anything of the sort, so while I was pretty sure it was not my fault, I was shaking the entire time I was dealing with the men to fill out the paperwork	«I was driving home after I got my dog from her sitter. My mom and dog were sitting in the passenger seat, **my dog likes to ride on the ground between my moms legs.** We came across quite a bit of traffic and were moving slowly towards our destination. As I move through this intersection, where cars on the right have a STOP sign (so they have to stop, and I have priority), this guy thats at least 60 years old, completely ignores the sign and advances towards us at quite a speed. Because there was traffic in front of me, I **could nothing but watch** as the car crashed into us, directly on the passengers side. **My mom was thrown to my side**, kept in place only by the seat belt, and **her leg was pretty badly hurt**, cause **she used her body to protect our doggy**. This experience is clear in my mind because after I got off **I had to have a fight with the other driver because he was incapable of acknowledging fault**».

Box 1. Example of a statement provided by participants during the task. Note: On the left side, a sample truthful statement from a participant telling when being involved in a car accident. On the right side, the same participant providing the deceptive statements about the same event following the given instructions. In bold the embedded lies identified by the participant.

Individual differences

We investigated individual differences in the absolute and standardized number of embedded lies, deceptiveness, and centrality scores. Regarding demographic factors (see also SM 2), we found a gender difference for the average deceptiveness scores (*diff* = 0.11 ± 0.05, *p* = .03, *d* = 0.14 [0.02, 0.26]), with females (*M* = 3.98, *SD* = 0.79) reporting higher values than males (*M* = 3.88, *SD* = 0.77). As for age, we found a small, significant positive correlation between age and deceptiveness (*rho* = 0.075, *S* = 172823388, *p* = .015).

Furthermore, we investigated the presence of subpopulations of liars by a cluster analysis of participants' scores in the four-factor lying profile questionnaire[49]. Lying profile scores were first adjusted for social desirability. The correction procedure employed a Generalized Linear Model approach to regress out the scores of each lying profile factor (i.e., LIE_Ability, LIE_Contextuality, LIE_Frequency, LIE_Negativity) for social desirability effects (i.e., SDE and IM). The adjusted scores were calculated using the *adjust* function from the *datawizard* package in Rstudio. We then followed the procedure in Makowski et al. (2021)[49] to cluster participants (see SM 3). To check if our dataset is appropriate for clustering, we computed the Hopkins' *H* statistic using the *check_clusterstructure* function from the *performance* package in Rstudio. Our dataset was deemed suitable for clustering, with Hopkins' *H* = 0.25 (i.e., a value for *H* lower than 0.25 indicates a clustering tendency at the 90% confidence level[64]). The method agreement procedure supported the existence of two clusters, as indicated by ten methods out of 29 (34.48%). After applying the k-means clustering algorithm, the two clusters accounted for 31.97% of the total variance of the original data. The first cluster (44.72% of the sample) was characterized by participants with very low reported lying ability, low levels of frequency and contextuality, and strong negative attitudes towards lying; the second cluster (55.28% of the sample) was characterized by people with higher levels of contextuality and frequency of deception, very high levels of ability and low levels of negative attitudes

	Embedded lies		
	M	***SD***	**Median**
Words	46.27	42.23	35
Absolute no. of embedded lies	5.03	3.35	4
Standardized no. of embedded lies	0.32	0.20	0.29
Deceptiveness	3.94	0.79	4
Centrality	3.55	0.82	3.59

Table 3. Descriptive statistics of participants' responses in variables associated with embedded Lies (M, SD, Median).

towards lying (Fig. 3). Following the original work[49], we labelled the first cluster as the *virtuous* and the second as the *trickster* cluster. To test the validity of this two-cluster solution, we trained a logistic regression that used, as features, the adjusted scores of the four scales of the lying profile questionnaire and, as a predicted variable, the labels obtained from the cluster analyses (as in[63]). We obtained an almost perfect classification (accuracy = 0.99). This result supported the validity of our two-cluster solution, confirming that the labels associated with each participant were not randomly assigned but actually reflected an inherently different pattern of responding. However, no significant differences were found in any dependent variable in the two groups (see Table 6S in SM 3).

Textual analysis of narrative data

Tables 4 and 5 suggest that a few linguistic indicators were significantly indicative of deception, albeit often with small effect sizes. LIWC variables associated with deceptive statements pertained mainly to using emotional and social words and references (i.e., social words, social references, pronouns and personal pronouns, social behaviour, language of status and leadership; Table 4). In contrast, LIWC features associated with truthfulness included mainly words associated with memory (i.e., remember, forget, remind) and numbers.

When we conducted the analysis by event, significant differences emerged for some LIWC variables by statement veracity for four events (i.e., Being hospitalized and undergoing surgery, Ending a long romantic relationship, Getting a speeding fine, and Taking the bus/train without the ticket).

For DeCLaRatiVE linguistic features (Table 5), only a few of them were significantly indicative of deception in five out of eleven events. When testing the whole dataset, the only significant features for deceptive statements were references to others, the number of words and number of syllables. In contrast, significant features for truthful statements were memory-related words and temporal details.

Finally, the *n*-gram differentiation analysis (Table 6) revealed how deceptive statements with embedded lies may appear very similar to their truthful counterparts, resulting in few or no significant differences in word usage. This result highlights the reasons why detecting embedded lies is a hard task.

Predictive modelling performance

We trained different machine learning and language models to distinguish deceptive statements with embedded lies from truthful ones. Table 7 shows that all models could classify statements better than the chance level (with p <.01 after running an exact binomial test), but the highest performance reached 64% accuracy after fine-tuning a Llama-3 model.

Exploratory explainability analysis

To add interpretations to the achieved performance, we conducted an explainability analysis on the Llama-3 and deception language model. We computed Spearman's rank correlations between the deceptive class probabilities and the absolute and standardized number of embedded lies, deceptiveness, and centrality scores (Table 7S in SM 4). There was a significant positive correlation between the class probability of deceptiveness and the absolute ($rho = 0.10$, $S = 170216978$, p <.01) and standardized number of embedded lies ($rho = 0.10$, $S = 170565831$, p =.001). For the deception language model, we found a significant positive correlation between the absolute number of embedded lies and class probability ($rho = 0.09$, $S = 171758230$, p =.004). Finally, only for the Llama-3 model, we found correct classifications having a significantly higher amount of absolute number of embedded lies ($M = 5.31$, $SD = 3.39$) compared to incorrect ones ($M = 4.43$, $SD = 2.83$), $d = 0.27$ [0.14, 0.40]. Similarly, a standardized number of embedded lies was significantly higher in correctly classified statements ($M = 0.34$, $SD = 0.21$) with respect to incorrect ones ($M = 0.29$, $SD = 0.19$), $d = 0.22$ [0.09, 0.35] (see Table 8S in SM 4). These findings suggest that the more a statement is fabricated, namely, the greater the number of embedded lies within an otherwise truthful statement, the higher the probability of a language model to accurately and confidently predict the class of that statement.

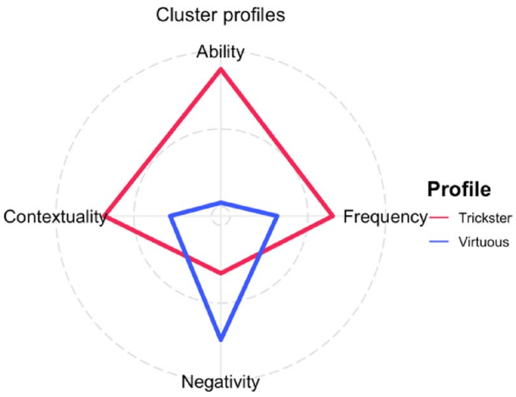

Fig. 3. Radar plot of the average values at the four lying profile factors in the trickster and virtuous cluster. The scores at the lying profile factors are corrected for Social Desirability.

Topic	LIWC feature	LIWC Interpretation	LIWC example words	Cohen's d	Adjusted CI	Direction
Overall	Social	Social words	Argue, boyfriend, chat	−0.22	−0.33, −0.11	D > T
	WC	Total word counts	-	−0.19	−0.30, −0.08	D > T
	Period	Number of periods	.	0.19	0.07, 0.29	T > D
	socrefs	Social references	you, we, he, she	−0.18	−0.29, −0.07	D > T
	shehe	Third singular personal pronouns	she, he, her, his	−0.18	−0.29, −0.07	D > T
	ppron	Personal pronouns	I, you, my, me	−0.17	−0.28, −0.06	D > T
	memory	Memory words	remember, forget, remind	0.17	0.06, 0.28	T > D
	socbehav	Social behavior words	said, love, say, care	−0.15	−0.25, −0.04	D > T
	male	Male references	he, his, him, man	−0.14	−0.25, −0.03	D > T
	number	Numbers	one, two, first, once	0.14	0.03, 0.25	T > D
	emo_anger	Emotion of anger	hate, mad, angry, frustr*	−0.14	−0.25, −0.03	D > T
	pronoun	Pronouns	I, you, that, it	−0.13	−0.24, −0.02	D > T
	det	Determiners	the, at, that, mine	−0.13	−0.23, −0.02	D > T
	Clout	Language of leadership, status	-	−0.11	−0.22, −0.001	D > T
Being hospitalized and undergoing surgery	Tone	Emotional tone	-	0.49	0.04, 0.94	T > D
	Period	Number of periods	.	0.48	0.03, 0.92	T > D
	WC	Total word count	-	−0.47	−0.92, −0.02	D > T
	power	Words of power	own, order, allow, power	−0.45	−0.89, −0.002	D > T
Ending a long romantic relationship	emo_anger	Emotion of anger	hate, mad, angry, frustr*	−0.41	−0.71, −0.12	D > T
	conflict	Conflict words	fight, kill, killed, attack	−0.36	−0.66, −0.06	D > T
	ppron	Personal pronouns	I, you, my, me	−0.34	−0.63, −0.04	D > T
	socbehav	Social behavior words	said, love, say, care	−0.32	−0.61, −0.02	D > T
Getting a speeding fine	Social	Social words	Argue, boyfriend, chat	−0.75	−1.26, −0.24	D > T
	socrefs	Social references	you, we, he, she	−0.70	−1.20, −0.19	D > T
	shehe	Third singular personal pronouns	she, he, her, his	−0.68	−1.18, −0.18	D > T
	Clout	Language of leadership, status	-	−0.56	−1.04, −0.07	D > T
Taking the bus/train without the ticket	Social	Social words	Argue, boyfriend, chat	−0.65	−1.00, −0.30	D > T
	shehe	Third singular personal pronouns	she, he, her, his	−0.57	−0.92, −0.23	D > T
	socrefs	Social references	you, we, he, she	−0.56	−0.90, −0.22	D > T
	male	Male references	he, his, him, man	−0.54	−0.88, −0.20	D > T
	socbehav	Social behavior words	said, love, say, care	−0.50	−0.84, −0.16	D > T
	comm	Communication words	said, say, tell, thank*	−0.46	−0.79, −0.13	D > T
	ppron	Personal pronouns	I, you, my, me	−0.41	−0.74, −0.08	D > T
	pronoun	Pronouns	I, you, that, it	−0.41	−0.74, −0.07	D > T
	Cognition	Words of Cognition	know, think, but, if	0.37	0.04, 0.70	T > D
	WC	Total word count	-	−0.35	−0.68, −0.03	D > T
	tentat	Words of tentativeness	if, or, any, something	0.34	0.01, 0.67	T > D
	visual	Visual words	see, lool, eye*, saw	−0.33	−0.65, −0.002	D > T

Table 4. Effect sizes (and CIs) of significant LIWC features for the entire dataset and specific events. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. Linguistic features are sorted by the absolute value of the effect size magnitude for each event. For the direction of the effect, T = truthful and D = deceptive.

Discussion

Moving forward on embedded lies

In this paper, we sought to spark renewed research interest in verbal deception detection and to move the dial towards embedded lies. With this aim, we presented a dataset of 2084 statements (i.e., truthful vs. deceptive with embedded lies) about eleven categories of autobiographical events deemed relevant for lying. We focused on autobiographical memories because of their relevance in forensic contexts, where the credibility of witnesses' and suspects' statements is assessed and often centred on autobiographical events. Additionally, this new dataset was collected in a within-subjects design, providing data at the statement and the individual level. Specifically, it provides granular information on the statement level, including annotations and ratings of embedded lies and memory-related measures about each event (e.g., how in the past, how frequently it is remembered, how important it is, etc.) and demographic data and personality-related measures, such as attitudes towards lying and social desirability, at the individual level. We believe this resource might be valuable in fostering psychological research on linguistic, contextual, and individual differences associated with embedded lies.

Topic	DeCLaRatiVE feature	Cohen's d	Adjusted CI	Direction
Overall	Other-reference	−0.19	−0.29, −0.10	D > T
	num_syllables	−0.19	−0.29, −0.09	D > T
	num_words	−0.19	−0.28, −0.09	D > T
	memory	0.17	0.07, 0.27	T > D
	Temporal details	0.10	0.0004, 0.19	T > D
A job interview for your dream job	memory	0.25	0.003, 0.51	T > D
Being hospitalized and undergoing surgery	Tone	0.49	0.09, 0.88	T > D
	num_words	−0.47	−0.87, −0.08	D > T
	num_syllables	−0.47	−0.86, −0.08	D > T
Getting a speeding fine	Other-reference	−0.66	−1.10, −0.22	D > T
Missing a deadline at work because of bad organization	Other-reference	−0.35	−0.67, −0.02	D > T
	memory	0.33	0.006, 0.66	T > D
Taking the bus without the train ticket	Other-reference	−0.52	−0.82, −0.22	T > D
	Cognition	0.37	0.08, 0.66	T > D
	num_words	−0.35	−0.64, −0.06	D > T
	num_syllables	−0.35	−0.64, −0.07	D > T

Table 5. Effect sizes (and CIs) of significant DeCLaRatiVE features for the entire dataset and specific events. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. Linguistic features are sorted by the absolute value of the effect size magnitude for each event. For the direction of the effect, T = truthful and D = deceptive.

Event	*n*-gram	*r*	Adjusted CI	Direction
Taking the bus/train without the train ticket	tell	−0.20	−0.37, −0.02	D > T
	ticket	−0.18	−0.23, −0.14	D > T
	time	−0.14	−0.26, −0.01	D > T
Ending a long romantic relationship	relationship	−0.07	0.001, 0.13	T > D
Missing a deadline at work because of bad organisation	time	0.10	0.004, 0.20	T > D
Cheating on your partner	feel	0.33	0.06, 0.60	T > D
Being hospitalized and undergoing surgery	pain	−0.22	−0.38, −0.06	D > T
	surgery	−0.14	−0.22, −0.05	D > T
Getting fired	fire	0.25	0.09, 0.40	T > D
Getting a speeding fine	speed	0.10	0.005, 0.20	T > D
Cheating on an exam	study	−0.28	−0.43, −0.13	D > T
	answer	0.19	0.01, 0.36	T > D
Causing a car accident	drive	−0.21	−0.38, −0.03	D > T

Table 6. Effect sizes (r) and CIs of significant n-grams for specific events after using the n-grams differentiation test. Confidence intervals are adjusted for multiple comparisons using Bonferroni correction. *N*-grams are sorted by effect size after comparing truthful and deceptive statements for each event. For the direction of the effect, T = truthful and D = deceptive.

The nature of embedded lies

Our findings suggest that participants used, on average, five embedded lies in their statements to achieve a predefined deception goal. About 1/3 of the length of deceptive statements were embedded lies. Similar figures are reported elsewhere for embedded lies in faked opinions about friends (37%)[2]. As for the source of embedded lies, most participants relied, whether directly or indirectly, on their personal experiences (46.27%), while a smaller percentage used their imagination (33.86%) or drew from others' experiences (14.95%). This finding supports the notion that liars often integrate elements of truth into their lies to enhance plausibility, making the detection of deception more difficult[2]. A realistic deceptive statement (i.e., one with embedded lies rather than a full-blown deceptive narrative) can thus be typified as one that consists of about 2/3 of truthful information and 1/3 of embedded lies, which are most likely to be derived from personal experience.

Individual differences in embedded lies

We further investigated individual differences in the nature of embedded lies. We found gender playing a role in how individuals self-rated the deceptiveness of their embedded lies, with females scoring higher in deceptiveness than males, albeit with small effect sizes. Age also played a role, with older participants being more openly deceptive in their statements.

		Truthful			Deceptive		
Model	**Accuracy**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BOW + RF	0.55 (0.03)	0.55 (0.03)	0.53 (0.04)	0.54 (0.03)	0.55 (0.03)	0.56 (0.04)	0.55 (0.03)
LIWC + RF	0.57 (0.03)	0.58 (0.03)	0.55 (0.03)	0.57 (0.03)	0.57 (0.03)	0.60 (0.05)	0.58 (0.04)
DeCLaRatiVE + RF	0.56 (0.02)	0.56 (0.02)	0.55 (0.06)	0.55 (0.04)	0.56 (0.02)	0.56 (0.06)	0.56 (0.03)
GPT-embeddings + RF	0.62 (0.03)	0.62 (0.03)	0.62 (0.05)	0.62 (0.03)	0.62 (0.03)	0.62 (0.05)	0.62 (0.03)
distilBERT	0.60 (0.02)	0.64 (0.05)	0.51 (0.19)	0.55 (0.10)	0.60 (0.05)	0.69 (0.16)	0.63 (0.05)
Fine-tuned FLAN-T5 base	0.60 (0.02)	0.60 (0.06)	0.57 (0.03)	0.59 (0.03)	0.59 (0.04)	0.63 (0.04)	0.61 (0.01)
Fine-tuned Llama-3-8B	**0.64 (0.04)**	**0.67 (0.05)**	**0.55 (0.13)**	**0.60 (0.08)**	**0.62 (0.05)**	**0.73 (0.10)**	**0.67 (0.05)**
Deception language model	0.56	0.54	0.76	0.63	0.60	0.35	0.44

Table 7. Classification performance of predictive models. The values refer to the average performance after performing cross-validation. In brackets, the standard deviation is reported. The deception language model was only employed to predict the class in our dataset; therefore, no cross-validation was performed. All models were significantly better than the chance level with p <.01. In bold is the performance of the best model. BOW = bag of words. RF = random forest. LIWC = Linguistic inquiry and word count.

In terms of lying attitude, the results of the cluster analysis were slightly different from the original paper[49]. We identified only two, rather than three, clusters of liars that resembled the original *virtuous* and *trickster* clusters. Specifically, the virtuous cluster was mainly characterised by a strong aversion to deception, while the tricksters tended to lie more frequently, to perceive themselves as good liars, and to adapt their lying behaviour to the context. However, despite this clear distinction, no significant differences were reflected in their behaviour and, specifically, in the absolute and standardised number of embedded lies, as well as in their deceptiveness and centrality scores. A possible explanation for why the difference in the lying attitude was not reflected in the lying behaviour (i.e., in the number of embedded lies) might be that all participants were instructed to write the statement deceptively by adding embedded lies, and this might have reduced the variability in their responses.

Textual properties of embedded lies

In addition to individual differences among liars, we examined linguistic properties of embedded lies by leveraging automated NLP techniques. Linguistic analysis using psycholinguistic variables and a deception-specific set of variables (DeCLaRatiVE) revealed few differences between truthful statements and those with embedded lies, with small effect sizes. Deceptive statements contained a larger proportion of social references, while truthful statements tended to include more references to memory processes. Similarly, the DeCLaRatiVE analysis suggested that deceptive statements contained more references to other people, which is in line with the distancing framework[52] and a higher number of words and syllables, which may reflect the experimental instructions that encouraged participants to add more details to achieve their deceptive goals. Conversely, truthful statements contained more memory-related words and temporal details, which is more in line with the Reality Monitoring framework[4,5].

When we zoomed in on the event level, we found significant differences in LIWC variables only in four out of eleven events and in DeCLaRatiVE variables in five out of eleven events. Altogether, these findings suggest that while there are some discernible differences between truthful and deceptive statements, these differences are often subtle and context-dependent. This is also in line with previous studies showing that truthful statements do not necessarily contain more details than embedded lies[2,33].

A term frequency analysis of n-grams underscored the difficulty of detecting deception through word usage when embedded lies are involved. In nine out of eleven events, we found negligible effects, with only one or two significant *n*-grams per event (e.g., "pain" and "surgery" as significant n-grams in deceptive statements for the event "Being hospitalized and undergoing surgery") and with small effect size, highlighting the subtle nature of embedded lies. This supports previous findings that verbal detection remains challenging, particularly when lies are carefully embedded within otherwise truthful narratives[2,68]. In addition, this overlap can be attributed to the within-subject design employed for this study, which eliminated any potential linguistic confounders derived from having different participants write about the same task under two conditions (honest vs. deceptive), typical of between-subjects studies.

Detecting embedded lies

By collecting a dataset that was sufficiently large to perform predictive modelling, we resorted to simpler supervised approaches based on machine learning models trained on extracted features and on state-of-the-art language models to classify statements as completely truthful or with embedded lies. The results showed that embedded lies present a significant challenge for deception detection due to their incorporation of truthful elements. Specifically, the highest performance of a language model with competitive capabilities (Llama-3.1-8B)[61], which we fine-tuned for this specific task, reached 64% accuracy. The result from our Llama model was in line with commonly reported performances in previous research[30,65–67]. Notably, a language model published in a previous study[22] - with a reported accuracy of 79.31% in detecting fabricated statements across different contexts - dropped to 56% accuracy when applied to our study. An explanation for that drop could be attributed to overfitting. Related works showed, in fact, deception classifiers dropping remarkably when tested on new samples[15]. However, we argue this was not the case. In the original study, the detection rate (i.e., the recall) for truthful (81%) and deceptive statements (78%) was balanced. In contrast, in our study, the deception language model showed a recall of 76% for truthful statements, similar to that of the original study, but a remarkable drop to 35% specifically for deceptive statements. This drop indicates that the struggle was mainly in the detection of embedded lies, which were often misclassified as truthful statements (here: 65% of embedded lies were misclassified as truthful, vs. 22% in the original study). If it were a matter of overfitting, we would have also expected a remarkable drop in the recall of truthful statements. However, this decline was not observed, which indicates that while the deception model was able to resort to what was learnt during the training phase to classify new samples of truthful statements correctly, it was unable to do so for the deceptive ones. We argue this was attributed to the fact that the deception involved was different (here: embedded lies vs. fabrication in the original study). Moreover, the explainability analyses on the Llama-3 and deception language model provided further evidence for the notion that the more nuanced the embedded lies are, the harder they are to detect.

Finally, when employing other common approaches, typically employed to detect deception (i.e., ML models trained on BOW representation, LIWC features, and embeddings), performance was significantly better than chance – albeit reaching just 55–62% accuracy. Other fine-tuned language models (here: distilBERT and FLAN-T5 base) were no more effective in performing the task. Altogether, these findings indicate that the challenge in identifying embedded lies stems from their resemblance to truthful statements and, as the degree of fabrication increases, the classification process becomes more straightforward.

Limitations and future outlooks

Despite the study's aim to overcome known limitations related to deception detection research (e.g., focus on fabrication, use of between-subjects designs, and small sample sizes), it comes with its own limitations in methodology and findings.

Regarding methodology, embedded lies were both self-reported and self-annotated by the participants, leading to subjective interpretations of what constitutes an embedded lie. This subjectivity could reduce the consistency and reliability of the data. In our analysis we standardized the number of embedded lies by computing the ratio of words in embedded lies to the total number of words in the deceptive statement to ensure that the results were not influenced by individual interpretations of what a unit of embedded lie was. We recommend future researchers adopting this or other forms of standardization (even during the data collection process) to ensure consistency. Second, while we recruited a Prolific sample that was sufficiently large to conduct meaningful computational analysis, these findings should be replicated with laboratory experiments where participants are in contact with the interviewer and can offer a longer verbal narrative, instead of a short written account. Additionally, previous research showed that more proactive interviewing techniques, such as the strategic use of evidence, the use of unexpected questions, or the Reality Interview (see[48] for an overview of these approaches), increase differences between truth-tellers and lie-tellers, enhancing deception detection rates. Therefore, further investigation on the detection of embedded lies using these interviewing approaches is needed, as they might promise higher accuracy rates. Third, while the dataset covers eleven distinct events, focusing the investigation on individual events, it may result in smaller sample sizes, limiting the statistical power and the ability to conduct predictive analyses within specific events. Finally - and in contrast to the study design employed by Markowitz[2] - we conceptualised truthful statements as entirely truthful, while deceptive statements were situated on a continuum ranging from embedded lies to completely fabricated statements. While it is reasonable that individuals may occasionally offer partial truths, it is also feasible to convey completely truthful statements. Consequently, we opted to narrow our focus of investigation by contrasting completely truthful statements with varying degrees of embedded lies. A potential avenue for future research could involve incorporating partial truths, as Markowitz did in his design[2], or alternatively, having three versions of the statement: truthful, embedded lies, and fully deceptive.

Regarding findings, we focused on predictive modelling, with the task being conceptualized as a classification task (i.e., whether a statement is truthful or contains embedded lies). However, future studies can go beyond this binary classification and conceptualize the task as a regression task where ML models quantify the extent of deception (e.g., the number of embedded lies) in a given statement. Additionally, future studies might focus on a sequence classification task to predict how and where lies are embedded within truthful narratives. Another obvious limitation is that the models tested were not compared with truth/lie judgments of untrained humans. Although meta-analytical evidence indicates that untrained judges perform close to the chance level[69], this finding is worth replication when lying involves embedded lies. Further, previous theoretical frameworks of deception and theories relying on manual coding, such as the use-the-best heuristic[70], the verifiability approach[6], as well as the role of complications, common knowledge details, or self-handicapping strategies[36], should be tested on this new dataset of deception to provide novel insights on what works on embedded lies.

Conclusion

In this paper, we presented a novel dataset as a resource to encourage research on embedded lies in verbal deception detection. The analysis of individual differences and linguistic properties, as well as the results from predictive modelling and explainability analysis, highlighted how the unique challenge in detecting embedded lies stems from their nuanced nature and resemblance to truthful statements.

Data availability

Data and scripts used to run the experiments are available at https://osf.io/jzrvh/.

Received: 16 January 2025; Accepted: 9 July 2025
Published online: 21 July 2025

References

1. Verigin, B. L., Meijer, E. H., Bogaard, G. & Vrij, A. Lie prevalence, lie characteristics and strategies of self-reported good liars. *PLoS One* **14** (12), e0225566 (2019).
2. Markowitz, D. M. Deconstructing deception: frequency, communicator characteristics, and linguistic features of embeddedness. *Appl. Cogn. Psychol.* **38**, e4215 (2024).
3. Amado, B. G., Arce, R., Fariña, F. & Vilariño, M. Criteria-Based content analysis (CBCA) reality criteria in adults: A meta-analytic review. *Int. J. Clin. Health Psychol.* **16**, 201–210 (2016).
4. Johnson, M. K. & Raye, C. L. Reality monitoring. *Psychol. Rev.* **88**(1), 67–85 (1981).
5. Sporer, S. L. Reality monitoring and detection of deception. in *The Detection of Deception in Forensic Contexts* 64–102Cambridge University Press, (2004). https://doi.org/10.1017/CBO9780511490071.004
6. Nahari, G., Vrij, A. & Fisher, R. P. Exploiting liars' verbal strategies by examining the verifiability of details. *Legal Criminol. Psychol.* **19**, 227–239 (2014).
7. Palena, N., Caso, L., Vrij, A. & Nahari, G. The verifiability approach: A meta-analysis. *J. Appl. Res. Mem. Cogn.* **10**, 155–166 (2021).
8. Verschuere, B., Bogaard, G. & Meijer, E. Discriminating deceptive from truthful statements using the verifiability approach: A meta-analysis. *Appl. Cogn. Psychol.* **35**, 374–384 (2021).
9. Gancedo, Y., Fariña, F., Seijo, D., Vilariño, M. & Arce, R. Reality monitoring: A Meta-analytical review for forensic practice. *Eur. J. Psychol. Appl. Leg. Context.* **13**, 99–110 (2021).
10. Fitzpatrick, E., Bachenko, J. & Fornaciari, T. *Automatic Detection of Verbal Deception* (Springer International Publishing, 2015). https://doi.org/10.1007/978-3-031-02158-9
11. Boyd, R. L., Ashokkumar, A., Seraj, S. & Pennebaker, J. W. *The Development and Psychometric Properties of LIWC-22.* (2022). https://www.liwc.app/
12. Mihalcea, R. & Strapparava, C. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. 309–312 Preprint at (2009). https://aclanthology.org/P09-2078
13. Ott, M., Choi, Y., Cardie, C. & Hancock, J. T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. 309–319 Preprint at (2011). https://aclanthology.org/P11-1032
14. Kleinberg, B., Mozes, M., Arntz, A. & Verschuere, B. Using named entities for Computer-Automated verbal deception detection. *J. Forensic Sci.* **63**, 714–723 (2018).
15. Kleinberg, B., van der Toolen, Y., Vrij, A., Arntz, A. & Verschuere, B. Automated verbal credibility assessment of intentions: the model statement technique and predictive modeling. *Appl. Cogn. Psychol.* **32**, 354–366 (2018).
16. Hauch, V., Blandón-Gitlin, I., Masip, J. & Sporer, S. L. Are computers effective lie detectors?? A Meta-Analysis of linguistic cues to deception. *Personality Social Psychol. Rev.* **19**, 307–342 (2015).
17. Constancio, A. S. et al. Deception detection with machine learning: A systematic review and statistical analysis. *PLoS One* **18**(2), e0281323 (2023).
18. Vaswani, A. et al. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017-December**, 5999–6009 (2017).
19. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. BERT: Pre-training of deep bidirectional Transformers for Language Understanding. *NAACL HLT 2019–2019 Conf. North. Am. Chapter Association Comput. Linguistics: Hum. Lang. Technol. - Proc. Conf.* **1**, 4171–4186 (2018).
20. Fornaciari, T. & Poesio, M. DeCour: a corpus of DEceptive statements in Italian COURts. 1585–1590 Preprint at (2012). http://www.lrec-conf.org/proceedings/lrec2012/pdf/377_Paper.pdf
21. Chung, H. W. et al. Scaling Instruction-Finetuned Language Models. Preprint at (2022). https://doi.org/10.48550/arXiv.2210.11416
22. Loconte, R., Russo, R., Capuozzo, P., Pietrini, P. & Sartori, G. Verbal lie detection using large Language models. *Sci. Rep.* **13**, 22849 (2023).
23. Wang, G., Chen, H. & Atabakhsh, H. Criminal identity deception and deception detection in law enforcement. *Group. Decis. Negot.* **13**, 111–127 (2004).
24. Bell, K. L. & DePaulo, B. M. Liking and lying. *Basic. Appl. Soc. Psych.* **18**, 243–266 (1996).
25. Leins, D. A., Zimmerman, L. A. & Polander, E. N. Observers' real-time sensitivity to deception in naturalistic interviews. *J. Police Crim Psychol.* **32**, 319–330 (2017).
26. DePaulo, B. M. et al. Cues to deception. *Psychol. Bull.* **129**, 74–118 (2003).
27. Hartwig, M., Granhag, P. A. & Strömwall, L. A. Guilty and innocent suspects' strategies during Police interrogations. *Psychol. Crime. Law.* **13**, 213–227 (2007).
28. Leins, D. A., Fisher, R. P. & Ross, S. J. Exploring liars' strategies for creating deceptive reports. *Legal Criminol. Psychol.* **18**, 141–151 (2013).
29. Vrij, A. et al. Increasing cognitive load to facilitate lie detection: the benefit of recalling an event in reverse order. *Law Hum. Behav.* **32**, 253–265 (2008).
30. Kleinberg, B. & Verschuere, B. How humans impair automated deception detection performance. *Acta Psychol. (Amst).* **213**, 103250 (2021).
31. Sap, M. et al. Quantifying the narrative flow of imagined versus autobiographical stories. *Proc Natl. Acad. Sci. USA* **119**(45), e2211715119 (2023).
32. Monaro, M., Maldera, S., Scarpazza, C., Sartori, G. & Navarin, N. Detecting deception through facial expressions in a dataset of videotaped interviews: A comparison between human judges and machine learning models. *Comput. Hum. Behav.* **127**, 107063 (2022).
33. Verigin, B. L., Meijer, E. H., Vrij, A. & Zauzig, L. The interaction of truthful and deceptive information. *Psychol. Crime. Law.* **26**, 367–383 (2020).
34. Verigin, B. L., Meijer, E. H. & Vrij, A. A within-statement baseline comparison for detecting Lies. *Psychiatry Psychol. Law.* **28**, 94–103 (2021).
35. Vrij, A. Baselining as a lie detection method. *Appl. Cogn. Psychol.* **30**, 1112–1119 (2016).
36. Caso, L., Cavagnis, L., Vrij, A. & Palena, N. Cues to deception: can complications, common knowledge details, and self-handicapping strategies discriminate between truths, embedded Lies and outright Lies in an Italian-speaking sample? *Front Psychol.* **14**, 1128194 (2023).
37. Levine, T. R., Serota, K. B., Carey, F. & Messer, D. Teenagers lie a lot: A further investigation into the prevalence of lying. *Communication Res. Rep.* **30**, 211–220 (2013).
38. Serota, K. B. & Levine, T. R. A few prolific liars: variation in the prevalence of lying. *J. Lang. Soc. Psychol.* **34**, 138–157 (2015).
39. Serota, K. B., Levine, T. R. & Boster, F. J. The prevalence of lying in america: three studies of Self-Reported Lies. *Hum. Commun. Res.* **36**, 2–25 (2010).
40. Hart, C. L., Lemon, R., Curtis, D. A. & Griffith, J. D. Personality traits associated with various forms of lying. *Psychol. Stud. (Mysore).* **65**, 239–246 (2020).
41. Kashy, D. A. & DePaulo, B. M. Who Lies?. *J. Pers. Soc. Psychol.* **70**, 1037–1051 (1996).
42. Weiss, B. & Feldman, R. S. Looking good and lying to do it: deception as an impression management strategy in job interviews. *J. Appl. Soc. Psychol.* **36**, 1070–1086 (2006).
43. Jones, D. N. & Paulhus, D. L. Duplicity among the dark triad: three faces of deceit. *J. Pers. Soc. Psychol.* **113**, 329–342 (2017).
44. Halevy, R., Shalvi, S. & Verschuere, B. Being honest about dishonesty: correlating self-reports and actual lying. *Hum. Commun. Res.* **40**, 54–72 (2014).

45. Semrad, M., Scott-Parker, B. & Nagel, M. Personality traits of a good liar: A systematic review of the literature. *Pers. Ind. Diff.* **147**, 306–316 (2019).
46. Paulhus, D. L. & Williams, K. M. The dark triad of personality: narcissism, machiavellianism, and psychopathy. *J. Res. Pers.* **36**, 556–563 (2002).
47. Kleinberg, B., Arntz, A. & Verschuere, B. Being accurate about accuracy in verbal deception detection. *PLoS One*. **14**, e0220228 (2019).
48. Vrij, A. et al. Verbal lie detection: its past, present and future. *Brain Sci.* **12**(12), 1644 (2022).
49. Makowski, D., Pham, T., Lau, Z. J., Raine, A. & Chen, S. H. A. The structure of deception: validation of the lying profile questionnaire. *Curr. Psychol.* **42**, 4001–4016 (2023).
50. Hart, C. M., Ritchie, T. D., Hepper, E. G. & Gebauer, J. E. The balanced inventory of desirable responding short form (BIDR-16). *Sage Open* **5**(4), 2158244015621113 (2015).
51. Pennebaker, J. W., Booth, R. J., Boyd, R. L. & Francis, M. E. Linguistic Inquiry and Word Count: LIWC2015 Operator's Manual, (2015). https://liwc.app/static/documents/LIWC2015%20Manual%20-%20Operation.pdf
52. Newman, M. L., Pennebaker, J. W., Berry, D. S. & Richards, J. M. Lying words: predicting deception from linguistic styles. *Pers. Soc. Psy Bullettin*. **29**, 665–675 (2003).
53. Vrij, A., Fisher, R. P. & Blank, H. A cognitive approach to lie detection: A meta-analysis. *Legal Criminol. Psychol.* **22**, 1–21 (2017).
54. Monaro, M. et al. Covert lie detection using keyboard dynamics. *Scientific Reports 2018 8:1* 8, 1–10 (2018).
55. Sporer, S. L. The less travelled road to truth: verbal cues in deception detection in accounts of fabricated and self-experienced events. *Appl. Cogn. Psychol.* **11**, 373–397 (1997).
56. Nahari, G., Vrij, A. & Fisher, R. P. The verifiability approach: countermeasures facilitate its ability to discriminate between truths and Lies. *Appl. Cogn. Psychol.* **28**, 122–128 (2014).
57. Brysbaert, M., Warriner, A. B. & Kuperman, V. Concreteness ratings for 40 thousand generally known english word lemmas. *Behav. Res. Methods*. **46**, 904–911 (2014).
58. Mozes, M., van der Vegt, I. & Kleinberg, B. A repeated-measures study on emotional responses after a year in the pandemic. *Scientific Reports 2021 11:1* 11, 1–11 (2021).
59. Ignatow, G. & Mihalcea, R. *Text Mining: A Guidebook for the Social Sciences. Text Mining: A Guidebook for the Social Sciences* (SAGE Publications, Inc, 2017). https://doi.org/10.4135/9781483399782
60. Sanh, V., Debut, L., Chaumond, J. & Wolf, T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2019).
61. Grattafiori, A. et al. The Llama 3 Herd of Models. Preprint at (2024). https://doi.org/10.48550/arXiv.2407.21783
62. Moore, J. H. Bootstrapping, permutation testing and the method of Surrogatedata. *Phys. Med. Biol.* **44**, L11 (1999).
63. Bambini, V. et al. Deconstructing heterogeneity in schizophrenia through language: a semi-automated linguistic analysis and data-driven clustering approach. *Schizophrenia 2022 8:1* 8, 1–12 (2022).
64. Lawson, R. G. & Jurs, P. C. New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* **30**, 36–41 (1990).
65. Rubin, V. L. & Conroy, N. Discerning truth from deception: human judgments and automation efforts. *First Monday* **17**(5), (2012).
66. Rubin, V. L. & Conroy, N. J. Challenges in automated deception detection in computer-mediated communication. *Proceedings of the American Society for Information Science and Technology* 48, 1–4 (2011).
67. Fornaciari, T. & Poesio, M. On the Use of Homogenous Sets of Subjects in Deceptive Language Analysis. 39–47 Preprint at (2012). https://aclanthology.org/W12-0406
68. Vrij, A. *Detecting Lies and Deceit: Pitfalls and Opportunities* (Wiley, 2008).
69. Bond, C. F. Jr & DePaulo, B. M. Accuracy of deception judgments. *Personality Social Psychol. Rev.* **10** (3), 214–234 (2006).
70. Verschuere, B. et al. The use-the-best heuristic facilitates deception detection. *Nature Human Behaviour 7:5* 7, 718–728 (2023).

Author contributions

Authors' contributionAll the authors worked on the conceptualization and investigation of the topic and decided on the most appropriate methods. R.L. worked on data curation, formal analysis and wrote the first draft of the paper. B.K. supervised the work and reviewed and edited the paper.

Funding

The authors received no financial support for the research, authorship, and publication of this article.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-11327-w.

Correspondence and requests for materials should be addressed to R.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.