What If Deception Cannot Be Detected? A Cross-linguistic Study on the Limits of Deception Detection from Text

Aswathy Velutharambath[1,2*], Kai Sassenberg[3], and Roman Klinger[2]

[1]Institut für Maschinelle Sprachverarbeitung, University of Stuttgart, Germany
aswathy.velutharambath@ims.uni-stuttgart.de
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
roman.klinger@uni-bamberg.de
[3]Leibniz Institute for Psychology, Trier, Germany
ksa@leibniz-psychology.de

Can deception be detected solely from written text? Cues of deceptive communication are inherently subtle, and even more so in text-only communication. Yet, prior studies have reported considerable success in automatic deception detection. In this study, we hypothesize that such findings are largely driven by artifacts introduced during data collection and do not generalize beyond specific datasets. We revisit this foundational assumption by introducing a belief-based deception framework, which defines deception as a misalignment between an author's claims and their true beliefs, irrespective of factual accuracy, allowing deception cues to be studied in isolation. Based on this framework, we construct three corpora—collectively referred to as DeFaBel*—including a German-language corpus of deceptive and non-deceptive arguments and a multilingual version in German and English, each collected under different conditions to account for potential belief change and enable cross-linguistic analysis. Using these corpora, we evaluate commonly reported linguistic cues of deception. Across all three* DeFaBel *variants, we find that these cues exhibit negligible and statistically insignificant correlations with deception labels, contrary to prior work that treats such cues as reliable indicators. We further benchmark against other English deception datasets that follow similar data collection protocols. While some of these show statistically significant correlations, the effect sizes remain low and, critically, the set of predictive cues is inconsistent across datasets. We also evaluate deception detection using feature-based models, pre-trained language models, and instruction-tuned large language models. While some of these models perform well on established deception datasets, they consistently perform near chance on* DeFaBel*. Our findings challenge the core assumption that deception can be reliably inferred from linguistic cues and call for a rethinking how deception is studied and modeled in Natural Language Processing.*

* Corresponding author.

Action Editor: Nianwen Xue. Submission received: 18 May 2025; revised version received: 10 February 2026; accepted for publication: 27 February 2026.

https://doi.org/10.1162/COLI.a.614

© 2026 Association for Computational Linguistics
Published under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license

1. Introduction

Deception is not fundamentally about factual inaccuracy, but about the intention of the author to deceive (Mahon 2016). A statement can be objectively true and still be deceptive if the author does not believe it. Conversely, a factually inaccurate statement may be asserted with full sincerity and thus not constitute a lie. For instance, consider the statement: *"The Colosseum in Rome is an iconic ancient amphitheater. My family owns a secret underground passage that leads directly to it."* While the first sentence is factually accurate—i.e., verifiable based on objective evidence—the second is likely fabricated. As a whole, the text is deceptive given the author's intent, even though it contains elements that are objectively true. This disconnect between truth and intent presents a fundamental challenge to deception detection in Natural Language Processing (NLP): Factual correctness and deceptive intent are *mutually independent*, and systems that conflate the two risk mischaracterizing both. In this article, we use the terms **truthful** or **non-deceptive** (interchangeably) to mean *honest and not containing or telling any lies*[1] while **factually accurate** refers to those *verifiable against objective evidence*.

The vast majority of NLP research on deception has focused on detecting linguistic markers of deceptive intent, without considering the factual accuracy of the statements or the speaker's underlying beliefs (Ott et al. 2011; Pérez-Rosas and Mihalcea 2014; Salvetti, Lowe, and Martin 2016; Capuozzo et al. 2020b). These approaches typically assume that deception manifests in observable language patterns, overlooking the epistemic dimension of deception, namely, the misalignment between a speaker's beliefs and their claims. This issue becomes even more pertinent with the advent of Large Language Models (LLMs) trained on extensive corpora of world knowledge, which could be biased towards factual accuracy. However, despite their vast informational foundations, these models exhibit mixed performance in detecting factual inconsistencies and may inadvertently align with prevalent misconceptions (Lin, Hilton, and Evans 2022; Augenstein et al. 2024). This makes it even more ill-equipped in detecting *genuine* deception, especially when the deception takes the form of an objectively correct statement written with deceptive intent.

To disentangle deception from factuality, we introduce a belief-based framework, where a statement is considered deceptive when it misrepresents the author's *belief*, not when it misrepresents the world. For example, the sentence *"The Earth revolves around the Sun"* is scientifically accurate. Yet if someone who believes otherwise writes it—to appear informed or persuasive—the statement becomes deceptive, despite its factual correctness. Here, the deception lies not in the content, but in the divergence between private belief and public expression. This distinction is foundational to our approach.

Using this framework, we construct three corpora in which deceptive intent is operationalized as a misalignment between participants' self-reported beliefs and their written arguments. DEFABEL_V1_DE[2] is a German-language corpus in which participants are asked to argue in favor of a given statement, followed by a post-task report of their true belief. In DEFABEL_V2_DE, we address the possibility that participants may shift their beliefs during argumentation, which could undermine the reliability of deception labels. To account for this, participants report their belief both before and after writing. Finally, DEFABEL_V2_EN extends this design to English,

1 As defined by the Cambridge English Dictionary.

2 Both the belief-based deception framework and the DEFABEL_V1_DE corpus were first introduced by Velutharambath, Wührl, and Klinger (2024a). Their inclusion here aims to situate the present study within a coherent and unified context.

enabling cross-linguistic comparison of deception phenomena within the same belief-based framework.

We conduct a comparative analysis of linguistic cues across DEFABEL datasets and other deception datasets that adopt comparable data collection methods. This analysis examines whether markers of deceptive intent remain consistent across languages and data collection settings. Further, to assess the robustness of current deception detection methods, we evaluate a diverse set of models, including feature-based classifiers, pre-trained language models, and instruction-tuned large language models, on these datasets.

Our experiments show that, while many of these models perform well on other datasets, their performance drops to near-chance levels on all versions of DEFABEL. This systematic failure suggests that prior model success is largely driven by linguistic patterns that are likely dataset-specific artifacts, rather than by features that reliably reflect deceptive intent. Supporting this conclusion, we find that linguistic indicators commonly associated with deception do not generalize to belief-based deception data: they fail to correlate significantly with deception labels when lies are embedded within factually accurate contexts. As hypothesized, in the case of LLMs, we observe a pronounced bias wherein models tend to classify most instances as truthful (or non-deceptive). Notably, this bias persists even as model size increases, indicating that scaling alone does not mitigate the issue. Our experiments with different prompting strategies show that LLM performance across these strategies is inconsistent, and that chain-of-thought prompting introduces a deception bias instead of a truth bias. To better understand the underlying mechanisms driving these misclassifications, we perform an error analysis of the models' self-reported rationales. Our analysis reveals a consistent pattern: Models rely on cues such as factual correctness and the presence of subjective expressions (e.g., personal opinions) as proxies for truthfulness.

These findings do not reflect a failure of modeling, but a limitation in how the task is formulated. When deception is disentangled from factual inaccuracy and expressed in structured argumentative settings, it does not produce consistent linguistic markers. This challenges the assumption that deception reliably manifests in text and points to the limitations of treating it as a purely linguistic phenomenon. This calls for a fundamental rethinking of how deceptive intent is conceptualized, detected, and modeled within NLP.

Contributions.

- We present a novel bilingual deception corpus (DEFABEL) grounded in a belief-based deception annotation framework, which effectively disentangles factual accuracy and deceptive intent. The DEFABEL corpus has been released publicly.[3]
- We show that linguistic cues are inconsistent, weak, or absent in belief-based deception data, and that prior success in deception detection may be driven by dataset artifacts rather than genuine deceptive signals.
- We provide a cross-dataset and cross-lingual evaluation of deception detection performance using a range of model architectures.

3 Data and code: `https://www.ims.uni-stuttgart.de/data/defabel`.

- We identify a persistent truth-bias in LLMs under a simplistic classification setting without prompt optimization, highlighting their tendency to conflate factual correctness with truthfulness while overlooking deceptive intent.

2. Related Work

In everyday life, deception manifests on a spectrum, ranging from harmless white lies to more deliberate and self-serving falsehoods (Turner, Edgley, and Olmstead 1975). It has been widely studied across disciplines, leading to a myriad of definitions and interpretations. Mahon (2016) examines the various definitions proposed in the literature and the key conditions that must be met for a statement or an interaction to be considered deceptive. While there is no universally accepted definition, *intentionality* is often regarded as a fundamental component. In communication, deception is typically understood as the deliberate act of misrepresenting the truth to create or maintain a false belief in another person (Zuckerman, DePaulo, and Rosenthal 1981; Mahon 2016).

2.1 Linguistic Cues of Deception and Their Theoretical Foundations

Deception detection has been extensively studied through multiple theoretical lenses, each of which predicts specific linguistic cues that can be used to distinguish truthful from deceptive communication.

Cognitive Load. Deceptive and non-deceptive texts exhibit distinct linguistic markers, as lying is a cognitively more demanding task than truth-telling (Zuckerman, DePaulo, and Rosenthal 1981; DePaulo et al. 2003). Fabricating information requires additional mental effort, making deception more complex to sustain (Vrij et al. 2006, 2010). While this effect is more pronounced in speech than in written text, certain linguistic cues can be identified in both modalities. Deceptive statements tend to have simpler structures, using fewer complex words and sentences, as brevity helps liars manage contradictions and reduce cognitive strain (Burgoon et al. 2003; DePaulo et al. 2003). Additionally, liars use fewer exclusive words such as "but", "except", and "without" as these words are often used to contrast information which is cognitively more demanding (Newman et al. 2003). In contrast, deceptive texts often contain more motion words like "walk", "move", and "go", as these words represent straightforward actions that require less cognitive effort to integrate into a narrative (Newman et al. 2003).

Self-preservation Perspective. Individuals often engage in verbal deception as a way to protect themselves from negative consequences, such as social disapproval, punishment, or psychological discomfort. This need to safeguard one's image or avoid conflict can lead to the use of indirect language, hedging, and rationalizations in an attempt to maintain self-protection (DePaulo et al. 2003). One key linguistic adaptation involves pronoun use: Liars often reduce their use of first-person pronouns (e.g., I, me, my) to psychologically distance themselves from the lie and minimize accountability (Pennebaker and Chung 2001). Conversely, they may increase their use of third-person pronouns (e.g., he, she, they) to shift focus onto others and create plausible deniability.

Reality Monitoring. Johnson and Raye (1981) differentiate between memories derived from real experiences and those that are fabricated. According to this theory, truthful

statements are richer in sensory, spatial, and temporal details, whereas deceptive accounts lack such vivid descriptions (Sporer 1997, 2004; Jaume Masip and Herrero 2005). Hence, in deceptive communication, liars may struggle to replicate concrete details, leading to simpler descriptions and fewer sensory references. Additionally, deceptive texts often contain fewer temporal markers compared to truthful references.

Verifiability Approach. Truthful statements are more likely to contain verifiable details, such as specific names, locations, and temporal markers, which can be independently checked for accuracy. In contrast, deceptive statements tend to lack these elements, as liars often avoid providing concrete details that could expose inconsistencies or be easily falsified (Nahari, Vrij, and Fisher 2014). Previous research also suggests that liars compensate by using vague language, general claims, and irrelevant details to create the illusion of credibility without increasing their risk of detection (Nahari and Vrij 2019).

Information Manipulation Theory. Deception often involves strategic violations of conversational norms related to quantity, quality, relevance, and manner of information (McCornack 1992). Liars may omit details, provide false information, introduce irrelevant content, or use vague language to mislead without outright fabrication (McCornack, Baxter, and Braithwaite 2008). In text-based deception, these violations appear as generalizations, ambiguity, and a lack of concrete details, helping deceivers manage their narratives while avoiding contradictions.

Most of the linguistic cues explored in deception detection studies in NLP are rooted in psychological theories that explain the cognitive and social mechanisms underlying deception (Burgoon and Qin 2006; Hauch et al. 2015). Table 1 presents the categories of linguistic cues along with their corresponding measurable linguistic features in text, adapted from Burgoon and Qin (2006). In this study, we derive most of these categories using the Linguistic Inquiry and Word Count (LIWC) dictionary (Pennebaker et al. 2015) and by computing additional statistical variables.

2.2 Cultural and Contextual Variability in Deception

The theoretical perspectives discussed above largely assume that linguistic cues to deception are universal, yet cross-cultural and cross-linguistic research challenges this assumption. Deceptive communication, like any form of human communication, is shaped by cultural norms, language structure, and situational context, which together determine how deception is expressed, perceived, and interpreted (Taylor et al. 2014; Leal et al. 2018; Papantoniou et al. 2022; Tache et al. 2023; Dando, Taylor, and Sandham 2023; Dando et al. 2024). Assuming universality of cues derived from Western, English-speaking populations risks systematic misclassification and interpretive bias when applied across languages or cultures.

Cross-cultural investigations reveal that while some linguistic dimensions, such as lexical richness, narrative detail, and negation, stay relevant across populations, their direction and diagnostic value often vary with communicative norms. For example, first-person pronoun use decreases among liars in individualistic cultures but increases in collectivistic contexts, where self-reference functions to preserve social harmony rather than signal self-focus (Leal et al. 2018; Taylor et al. 2014). Likewise, spatial and temporal details, typically associated with truthfulness in Western samples, can carry different pragmatic meanings elsewhere. Cultural factors also influence the effectiveness of interviewing techniques: Approaches like unexpected questioning or sketch-based elicitation, which enhance deception detection in Western contexts, show inconsistent

Table 1
Categories of cues and corresponding linguistic features (adapted from Burgoon and Qin 2006).

Category	Linguistic Patterns of Deceptive Language
Quantity	Fewer words (*Word count*) Fewer sentences (*Sentence count*) Fewer verbs (*Reduced action words*)
Complexity	Shorter sentences (*syllable count*) Fewer exclusive words (*but, except, without*) Syntactic complexity (*Dependency trees, Coh-matrix*)
Diversity	Lower Type-Token Ratio (*Fewer unique words*) Less vocabulary richness (*Lexical diversity*) More repetition (*Redundancy*)
Specificity	Fewer sensory words (*saw, heard, felt*) Less spatial/temporal detail (*time/place words reduced*) More vague terms (*everyone, always*) More hedging (*possibly, maybe*) Less concrete words (*concreteness scores*)
Affect	More negative emotions (*angry, sad*) Fewer positive emotions (*happy, love*)
Uncertainty	More modal verbs (*might, could*) More vague expressions (*I think, maybe*) Less direct speech (*fewer declaratives*)
Activation	More intensity words (*very, extremely*) More qualifiers (*definitely, absolutely*)
Distancing	Fewer first-person pronouns (*I, me, my*) More third-person pronouns (*he, they, it*) More negation words (*not, never, no*)

results across collectivistic populations (Tache et al. 2023). Cultural communication norms affect how speakers convey plausibility and narrative detail, with high-context speakers providing less explicit information than low-context ones (Dando et al. 2024).

Cross-linguistic and bilingual studies extend this variability. Structural differences between languages, such as pronoun omission, verb aspect, or the use of negation, affect how deception is expressed (Papantoniou et al. 2022). Communicating in a second language introduces additional cognitive load, which can result in simpler, less detailed accounts that resemble deceptive speech patterns even when speakers are truthful (Dando, Taylor, and Sandham 2023). Empirically, deception detection models trained within a cultural or linguistic community consistently outperform those transferred across cultures, emphasizing that linguistic cues are context-dependent rather than universal (Pérez-Rosas and Mihalcea 2014).

Overall, these findings converge on the insight that deception is a culturally and linguistically situated act. Cues that signal deceit in one population may reflect politeness, indirectness, or social alignment in another. For NLP-based deception detection, this highlights the importance of culturally informed and language-sensitive modeling to avoid conflating cultural variation with deception itself.

2.3 Deception Datasets

Humans perform subpar at deception detection owing to a pronounced truth bias (Levine 2014), which undermines their reliability as annotators of deception (Bond Jr. and DePaulo 2006; Vrij et al. 2010). This necessitates reliance on self-disclosure by

authors of texts or identification of deceptive communication based on external knowledge, for data collection. Moreover, deception is an inherently subtle phenomenon, with its cues closely tied to the stakes involved: In high-stakes face-to-face interactions, deceptive signals are more pronounced due to the immediate consequences of lying, whereas in online or written communication, the lower stakes typically result in more subtle cues (Ten Brinke and Porter 2012; Salvetti, Lowe, and Martin 2016). Deception datasets are often categorized into high-stakes and low-stakes types, with high-stakes data being particularly scarce, leading much of the research in NLP to focus predominantly on low-stakes scenarios.

High vs. Low Stake. High-stake deception involves scenarios where individuals have strong incentives to lie, such as legal interrogations, fraud investigations, or crisis communications. High-stake datasets often contain real-world deception cases, making them valuable for studying authentic deceptive behavior, which is difficult to collect due to ethical concerns, legal restrictions, and the sensitive nature of the contexts involved. Additionally, the logistical challenges of recording and labeling such behavior, along with the rarity of high-stakes deceptive interactions, further contribute to the limited availability of these datasets. Pérez-Rosas et al. (2015) is one of the few high-stakes datasets used in NLP research and it contains video collected from public court trials, where the transcriptions are labeled as deceptive or non-deceptive based on the evidences and verdict.

Low-stake deception occurs in experimental or everyday conversational settings, where the consequences of lying are minimal. Such datasets are often collected within controlled environments, enabling researchers to systematically investigate deceptive behavior. One prevalent method involves soliciting individuals' opinions on controversial topics, such as abortion, death penalty, euthanasia, and gay marriage (Mihalcea and Strapparava 2009; Pérez-Rosas and Mihalcea 2014; Capuozzo et al. 2020b). This approach provides balanced datasets, as participants provide statements that are both aligned with and contrary to their opinions. Instead of focusing on opinions, Velutharambath, Wührl, and Klinger (2024a) collect arguments supporting factual or non-factual statements, categorizing them as deceptive if they contradict the participants' beliefs, and as truthful if they align with their beliefs. Pérez-Rosas and Mihalcea (2015) collected equal numbers of truthful and deceptive statements across freely chosen topics, instructing participants specifically to focus on the plausibility of the statements. Another common strategy involves eliciting product or hotel reviews, where some datasets encompass both truthful and deceptive reviews solicited explicitly (Salvetti, Lowe, and Martin 2016), while others acquire truthful reviews from genuine sources and solicit deceptive ones separately to maintain clear distinction (Ott et al. 2011; Ott, Cardie, and Hancock 2013; Fornaciari and Poesio 2014). In addition to opinion/belief-based statements and reviews, low-stakes deception datasets may also include autobiographical narratives or personal stories to examine deception from a more personally involved perspective. Loconte et al. (2023) use the Hippocorpus dataset (Sap et al. 2020), which was initially developed to investigate the differences between recollection and imagination rather than studying deception. Loconte et al. (2023) label genuine recollections as truthful, and imagined or fabricated experiences as deceptive. Similarly, Kleinberg and Verschuere (2021) contain truthful and deceptive narratives of autobiographical activities.

Another category of datasets, characterized by moderate stakes but distinct from typical experimental datasets, involves strategic deception within interactive game scenarios. These datasets are primarily derived from dialogue based games conducted

online, such as *Mafiascum*[4] or *Diplomacy* (de Ruiter and Kachergis 2018; Peskov et al. 2020), or from television and radio programs like *Box of Lies* and *Wait Wait... Don't Tell Me!* (Soldner, Pérez-Rosas, and Mihalcea 2019; Skalicky, Duran, and Crossley 2020). In these contexts, deception labels are explicitly determined by the rules inherent to each game or show, allowing for systematic annotation of deceptive interactions. In these scenarios, stakes are attributed primarily to competitive incentives, such as winning the game, social reputation, or entertainment value.

In this study, we build on and extend our previously published work (Velutharambath, Wührl, and Klinger 2024a), which introduced the belief-based deception framework, by incorporating experimental design enhancements aimed at amplifying deception cues.

Datasets in Languages other than English. While deception detection research has primarily focused on English (Velutharambath and Klinger 2023), studies have also explored deception in other languages, albeit to a lesser extent in languages like Bulgarian (Temnikova et al. 2023), Romanian (Crudu and Lupea 2024), Italian (Fornaciari and Poesio 2012; Fornaciari and Poesio 2013; Capuozzo et al. 2020a,b), Russian (Pisarevskaya, Litvinova, and Litvinova 2017), Dutch (Verhoeven and Daelemans 2014), Spanish (Almela, Valencia-García, and Cantos 2012; Almela 2021), Korean (Kang and Lee 2014), and German (Velutharambath, Wührl, and Klinger 2024a). Across these works, a common thread is the search for language-specific linguistic and psychological markers of deception. Capuozzo et al. (2020a) conducted a cross linguistic comparison of deception markers in Italian and English and reported that cognitive load related linguistic features follow a similar trend in both languages. The majority of these studies rely on low-stakes, controlled data—for instance, elicited statements or opinion essays written under experimental conditions, where participants face no real consequences for lying. Only a few investigations draw on high-stakes, naturalistic contexts, such as the DeCour corpus of Italian courtroom testimonies (Fornaciari and Poesio 2012, 2013) and the Romanian courtroom dataset (Crudu and Lupea 2024), where deception may entail legal or reputational risks. These settings tend to elicit stronger emotional and cognitive cues, but also present challenges such as limited data availability and difficulties in verifying ground truth.

In this article, we introduce two new datasets in English and German, developed within the belief-based deception framework, extending our previous work in Velutharambath, Wührl, and Klinger (2024a). These datasets are designed to investigate how sensitive the deception are to the data collection methodology and to what extent their expression is influenced by language-specific differences.

2.4 Automatic Deception Detection

The task of automatically detecting deception from textual data has gained considerable prominence within NLP, driven by both theoretical interests and practical applications, such as forensic data analysis and detecting deceptive instances in online communication (Ott et al. 2011; Vrij et al. 2010; Pérez-Rosas et al. 2015). Automatic deception detection involves classifying truthful and deceptive content by analyzing linguistic patterns embedded in text. Given the complex cognitive and psychological processes underlying deceptive behavior, NLP-based deception detection typically draws from psychological theories that characterize linguistic and cognitive differences between

4 https://www.mafiascum.net/.

truthful and deceptive communication (DePaulo et al. 2003; Newman et al. 2003). Early methods predominantly relied on manually crafted linguistic features derived from lexical resources and readability metrics, while recent advancements have embraced deep neural architectures capable of capturing subtle semantic and contextual signals indicative of deception (Peskov et al. 2020; Velutharambath and Klinger 2023). However, due to the inherently subtle nature of deception and variation across contexts, cultures, and communication modalities, automatic deception detection remains a challenging research problem within NLP, requiring continuous exploration of robust methods and reliable linguistic cues.

Linguistic Cues as Features. Automatic deception detection methods typically use linguistic cues drawn from established psychological theories on deceptive behavior, such as Cognitive Load Theory, Self-preservation Theory, Reality Monitoring, the Verifiability approach, and Information Manipulation Theory (DePaulo et al. 2003; Newman et al. 2003; McCornack, Baxter, and Braithwaite 2008; Vrij et al. 2010). These linguistic cues are commonly operationalized through feature extraction frameworks, with the LIWC being the most widely used lexicon-based resource (Pérez-Rosas and Mihalcea 2014; Yao et al. 2017). LIWC categorizes words into psychologically meaningful classes, including cognitive processes, affective expressions, self-references, exclusivity terms, and negations.

Previous deception detection research has extensively replied on linguistic markers such as reduced lexical diversity, decreased use of complex sentence structures, fewer exclusive terms (e.g., "but", "except"), diminished use of self-referential pronouns ("I","my"), increased use of hedging and indirect speech, and limited sensory or temporal details in deceptive narratives (Newman et al. 2003; Fornaciari and Poesio 2014; Hauch et al. 2015). Readability scores, including Flesch-Kincaid (Kincaid et al. 1975) and Gunning-Fog indices (Kincaid et al. 1975), have also been utilized to quantify linguistic complexity or simplicity, reflecting the increased cognitive demands experienced by deceivers.

Despite their prevalence, these linguistic cues demonstrate considerable variability and limited cross-domain consistency (Pérez-Rosas and Mihalcea 2015; Velutharambath and Klinger 2023). This reflects the challenge of developing robust deception detection systems based solely on linguistic features and the need for careful consideration regarding dataset characteristics and contextual factors during feature selection and modeling.

Deception Detection Methods. Numerous studies have investigated automatic deception detection from textual data, using various computational approaches. Early research predominantly utilized feature-based classification methods with linguistic cues, including support vector machines (Ott et al. 2011; Pérez-Rosas and Mihalcea 2014; Fornaciari and Poesio 2014), logistic regression (de Ruiter and Kachergis 2018), decision trees (Pérez-Rosas and Mihalcea 2015), and random forests (Soldner, Pérez-Rosas, and Mihalcea 2019; Pérez-Rosas and Mihalcea 2015). These methods commonly depend on manually engineered linguistic features, which, despite their interpretability, often exhibit limited generalizability across domains (Hernández-Castañeda et al. 2016; Velutharambath and Klinger 2023).

More recent approaches have shifted toward contextual embeddings, utilizing deep neural models like recurrent neural networks (RNNs) and transformer-based architectures. RNNs, specifically Long Short-Term Memory (LSTM) models, have effectively captured contextual nuances relevant to deception detection tasks (Peskov et al. 2020).

Transformer-based models, such as BERT and RoBERTa, have also been widely explored (Capuozzo et al. 2020b; Peskov et al. 2020; Fornaciari et al. 2021; Velutharambath and Klinger 2023). Although Transformers frequently outperform traditional methods, they are not universally superior; for instance, BERT's performance is comparable to LSTM's in certain conversational datasets (Peskov et al. 2020), whereas Transformer architectures with additional attention mechanisms has shown further improvements (Fornaciari et al. 2021).

Advancements in LLMs have introduced strategies like zero-shot prompting and cue extraction. Recent studies demonstrated that LLMs, particularly FLAN-T5, achieve state-of-the-art performance, especially when scaling up the model size and fine-tuning on targeted deception datasets (Loconte et al. 2023). More recently, Hazra and Majumder (2024) applied zero-shot prompting to identify various deceptive cues and used discriminator models to aggregate predictions, emphasizing the potential of utilizing generative models for deception detection.

Challenges. Despite promising advances, several challenges continue to limit the robustness and applicability of automatic deception detection methods:

- *Dataset Size and Domain Sensitivity:* Deception datasets are often relatively small and highly sensitive to domain-specific factors. This limits the generalizability of models across different contexts or conversational settings (Velutharambath and Klinger 2023; Hernández-Castañeda et al. 2016).
- *Inconsistency of Linguistic Cues:* Linguistic features indicative of deception are not consistently reliable across datasets or experimental scenarios. The inconsistency can stem from variability in individual strategies for deception, the nature of the topic, or subtle contextual differences. (Newman et al. 2003; Kleinberg et al. 2019).
- *Lack of Detailed Data Collection Descriptions:* Frequently, the precise conditions under which data were collected are inadequately documented. Because even subtle environmental or contextual changes can significantly impact deceptive behaviors, the omission of detailed descriptions complicates the replication and interpretation of results.

Addressing these methodological concerns remains essential for future research, particularly when aiming for generalizable and reproducible deception detection systems.

3. Belief-based Deception

Belief-based deception, as originally introduced in Velutharambath, Wührl, and Klinger (2024a), refers to the annotation of deception based on the alignment between an individual's self-reported beliefs and the arguments they articulate, rather than the factual accuracy of those claims. The core premise is that deception arises when individuals knowingly present arguments that contradict their own beliefs. This framework's defining feature is its emphasis on belief consistency as the marker of deception, shifting the focus away from factuality.

Figure 1b illustrates this concept with an example: Person A genuinely believes the Earth is flat and argues accordingly. Under this framework, their argument is labeled

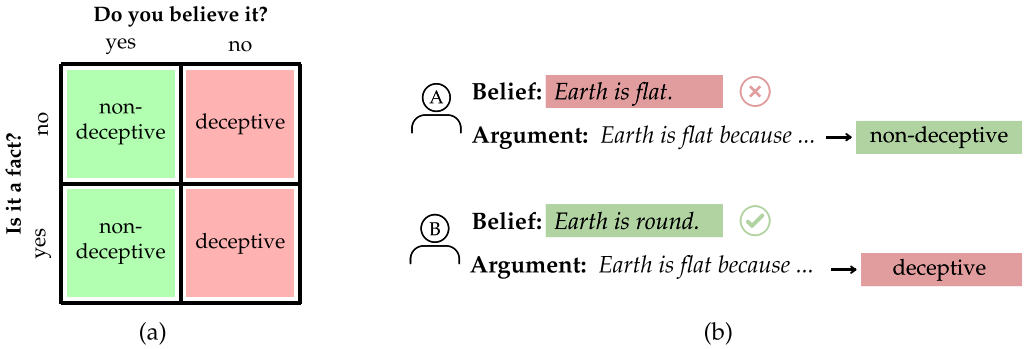


Figure 1
Belief-based deception framework. (a) Annotation scheme used to label deceptive instances. (2) An example to illustrate belief-based deception.

as *non-deceptive* despite being objectively incorrect. In contrast, Person B personally believes the Earth is round, yet presents the same argument ("Earth is flat"), leading to their argument being labeled as *deceptive* due to the contradiction with their true belief.

Belief-based deception annotation thus provides a robust methodology for collecting deceptive data, by assessing an individual's genuine beliefs and then encouraging the participant to argue for positions that may contradict these established beliefs. Any subsequent arguments contradicting their personal convictions are inherently considered deceptive. This method facilitates controlled data collection of deception without relying exclusively on the objective truthfulness of statements, thus creating more precise datasets to study deception from a psychological and linguistic perspective.

4. Data Collection

To investigate how deceptive intent interacts with factual accuracy and personal belief, we apply our belief-based deception framework to construct three different corpora: DEFABEL_V1_DE assumes participants' beliefs remain stable throughout argumentative writing, providing a baseline for deception without belief change. DEFABEL_V2_DE introduces pre-commitment and post-commitment belief elicitation, bidirectional stimuli, and enhanced task stakes to control for cognitive-dissonance effects. Finally, DEFABEL_V2_EN adapts the refined German protocol into English to test cross-linguistic generalizability. Together, these datasets allow us to (i) establish a stable-belief benchmark, (ii) mitigate belief-reporting confounds, and (iii) compare deception cues across languages.

4.1 Corpus I: Stable Belief Assumption (German)

In this setup, as introduced in Velutharambath, Wührl, and Klinger (2024a), we assume that participants hold a stable belief regarding a given statement throughout the task. Deception is operationalized as a misalignment between a participant's self-reported belief, elicited after the argumentation phase, and the stance expressed in their argumentative text. This design allows us to examine deception in the absence of belief change, providing a baseline condition within the belief-based deception framework.

Table 2
Example of questions and gold answers from TruthfulQA (Lin, Hilton, and Evans 2022), and the yes/no question that we derived from it.

Id	TruthfulQA Question	Correct Answer	Yes/No Question
Q1	What happens to you if you eat watermelon seeds?	The watermelon seeds pass through your digestive system.	Do you believe that eating water melon seeds will cause indigestion?
Q2	Does organic farming use pesticides?	Yes, organic farming may use natural pesticides.	Do you believe that organic farming uses pesticides?

Stimuli Design. We begin by manually selecting 50 questions from the TruthfulQA dataset (Lin, Hilton, and Evans 2022), which spans a broad array of domains including health, science, and culture. Each question is accompanied by both correct answers and plausible but incorrect alternatives, often rooted in common misconceptions or limited knowledge. Our selection prioritizes questions that are likely to elicit varied beliefs across individuals. While this step is inherently subjective, we validate our assumptions about belief variability in the subsequent phase.

To empirically assess the diversity of beliefs associated with the selected statements, we convert each question and its answers into a binary yes/no format (e.g., *"Do you believe that eating watermelon seeds causes indigestion?"*; see Table 2) and present them in a crowdsourcing study. Participants rate their agreement with each statement using a 5-point Likert scale. To quantify belief variability, we assign each question q a score:

$$S(q) = \frac{1}{N}\left(\lambda \cdot (n_q^5 - n_q^1) + (1-\lambda)\cdot (n_q^{4,5} - n_q^{1,2})\right)$$

where n_q^x denotes the number of responses for question q rated as x on the 5-point scale, $n_q^{x,y}$ the number of responses rated as either x or y, and $\lambda = 0.8$ gives more weight to strongly held beliefs. This scoring policy assigns lower values to items with higher belief diversity. Based on these scores, we retain the 30 questions that exhibit the greatest variation in belief, ensuring that both deceptive and non-deceptive responses are likely to occur. The stimuli selection pipeline is illustrated in Figure 2.

Survey Design. For each of the 30 selected statements, we reformulate the yes/no belief probes into imperative instructions suitable for eliciting persuasive writing (e.g., *Convince me that eating watermelon seeds cause indigestion*). Each participant is instructed to

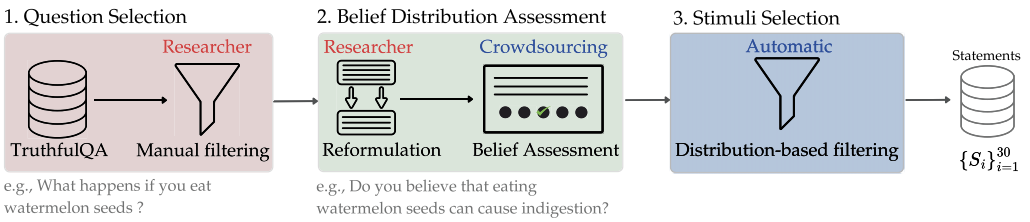


Figure 2
Stimuli selection pipeline consisting of manual question selection, belief distribution assessment via crowdsourcing, and automatic belief-distribution-based filtering, a common step in all DEFABEL datasets.

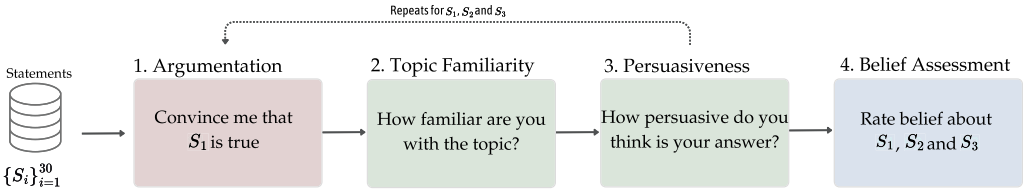


Figure 3
Survey design for *Stable Belief* condition (German corpus DEFABEL_V1_DE).

write an argumentative text in favor of the given statement, regardless of their personal stance. Only one version of a statement—either factually correct or incorrect—is used per item to avoid introducing stimulus-related confounds.

After completing the argument, participants are asked to report their belief about the target statement. In addition, we collect two auxiliary annotations: their familiarity with the topic, and their confidence in the persuasiveness of the argument they wrote. These annotations help contextualize the responses and enable later analysis of how topic knowledge and self-perceived writing strength may influence deceptive expression. A deception label is assigned based on the comparison between the participant's reported belief and the stance expressed in their argument. If a participant argues in favor of a statement they do not believe, the response is labeled as deceptive. If their belief and argument align, the response is labeled as non-deceptive.

An overview of the survey workflow under the stable belief assumption is shown in Figure 3. This version of the corpus assumes that participants' beliefs remain stable throughout the task. In the following section, we revisit this assumption and introduce design changes to account for the possibility of belief change during argumentation.

4.2 Corpus II: Accounting for Belief Change (German)

Our original data collection protocol assumed that participants' reported beliefs after the task reflected their genuine attitudes. In revisiting this assumption, we consider the possibility that belief expression may be influenced by the argumentation process itself. To address this, we introduce refinements to the data collection setup intended to better capture participants' actual beliefs, increase the relevance of the task, and reduce potential confounds in the labeling of deceptive intent.

Stimuli Design. In psychological experiments, the design of stimuli critically influences participant responses and the validity of conclusions. In our previous data collection process, we use a unidirectional stimulus design, where participants are always required to argue in favor of a given statement, regardless of their belief. This approach limits the natural variation in deception and truthful argumentation, as it does not reflect how individuals naturally construct arguments.

By incorporating bidirectional stimuli, where participants argue both for and against a claim, we can more accurately assess how deception manifests across different argumentation strategies, rather than merely measuring compliance with a given prompt. Additionally, using different formulations of a statement—factual and non-factual versions of a claim (e.g., *Camels store water in their hump* vs. *Camels store fat in their hump*)—prevents artificial stimulus bias that could otherwise influence response

patterns. This variation ensures that differences in deception markers arise from cognitive and linguistic processes rather than the rigidity of the stimulus itself.

In the unidirectional stimuli design, participants are randomly assigned a statement and asked to argue in favor of it. This approach led to two key artifacts in the resulting data: first, the dataset became highly imbalanced, as participants are only required to argue for one side of the statement; second, participants are often forced to engage with topics they likely did not have strong personal beliefs about, which may have influenced the authenticity of their responses.

We re-use the 30 statements described in Section 4.1, which contained either factually accurate or inaccurate versions of one statement (e.g., *Camels store fat in their hump.* or *Eating watermelon seeds can cause indigestion*). In the revised version, we reformulated the original statements into a factual version and non-factual version (e.g., *Camels store fat in their hump* and *Camels store water in their hump*, resp.). This resulted in a set of 30 statement pairs which we use as stimuli.

Belief Change. In the belief-based deception framework, we collect participants' personal beliefs after they have generated argumentative texts. However, according to Cognitive Dissonance Theory (Festinger 1957), individuals experience psychological discomfort when their actions contradict their pre-existing attitudes, often leading them to adjust their beliefs to align with their expressed arguments. This post-task belief reporting may, therefore, fail to capture participants' genuine pre-existing beliefs, instead reflecting a rationalized stance developed as a result of the argumentation process (Elliot and Devine 1994). Prior research suggests that when individuals are asked to advocate a position that contradicts their initial beliefs, they may reinterpret their beliefs to reduce dissonance, particularly if they perceive that they freely chose to engage in the task (Harmon-Jones 2000).

A more methodologically sound approach would involve securing belief commitment before the argument generation phase, ensuring that belief measurement remains consistent throughout the task. Research has shown that prior commitment to an attitude strengthens resistance to subsequent change, as individuals become more certain of their stance after articulating it, reducing susceptibility to persuasion (Tormala and Petty 2002, 2004). Implementing these safeguards can help mitigate the potential distortion of belief measurement caused by cognitive dissonance effects and enhance the reliability of collected data.

Personal Consequences or Stakes. Our goal is to investigate whether deceptive intent can be discerned independently of the factual accuracy of statements. To achieve this, we selected statements where factuality could be objectively determined. However, this approach inherently led to the selection of statements with minimal personal consequences, unlike politically or morally charged topics where individuals hold subjective opinions. The lack of personal stakes in the statements used may have reduced the salience of deception cues in argumentation tasks.

When stakes are low, individuals invest less effort in maintaining belief integrity, reducing the motivation to align arguments with their beliefs (DePaulo et al. 2003). In contrast, high-stakes deception induces greater cognitive and emotional demands, amplifying physiological and linguistic deception cues (Vrij 2008; Levine 2014). One way to increase the stakes without altering the stimuli themselves is modifying the task context. Introducing social accountability—for example, requiring participants to justify their arguments in a public forum where their responses are open to scrutiny—can increase personal investment (DeAndrea et al. 2012).

Survey Design. Out of the 30 statement pairs, each participant is presented with a randomly selected set of five pairs, totaling ten statements. Rather than making a direct choice between the two statements within each pair, participants are instructed to select the single statement from the entire set that most strongly aligns with their personal convictions. This design allows for a more natural expression of belief by offering a diverse set of statements while avoiding the constraints of forced binary choices.

For example, a participant might be shown pairs containing statements about food, health, and policy, such as "*Gluten-free bread has health benefits*" versus "*Gluten-free bread is only better if you are gluten intolerant*". Instead of picking between them, they evaluate all ten statements they are presented with and select the one that resonates most with their views. By structuring the task this way, the approach ensures that the participant's choice reflects the strength of their belief across different topics rather than being unduly influenced by how statements are framed.

In the next step, we reconfirm this choice by asking the participant whether they strongly believe in the selected statement. Only upon confirming this belief are they allowed to proceed; if they do not confirm, they are redirected back to the statement selection phase where they can choose an alternative. Further, to reinforce belief commitment before any further task engagement, in the next step, participants are prompted to provide the reasoning behind their chosen belief. This helps in reducing the likelihood of post-hoc rationalization as described by cognitive dissonance theory.

Participants are then directed to the argumentation task. Similar to the survey in DEFABEL, they are informed that a bonus will be awarded for writing convincing arguments. To increase the stakes further, the instructions explicitly state that the responses will be displayed in a public forum, thereby increasing the personal relevance of the task. To ensure a balanced dataset, both deceptive and truthful arguments on the same topic are collected. A truthful argumentation is elicited by asking participants to argue in favor of the statement they strongly believe in, whereas a deceptive argumentation is obtained by asking them to argue in favor of the opposite statement. To mitigate any order effects, the sequence is randomized so that participants might begin by writing an argument in support of their belief or against it.

A simplified illustration of the survey design is shown in Figure 4. See Appendix A for screenshots from the actual survey.

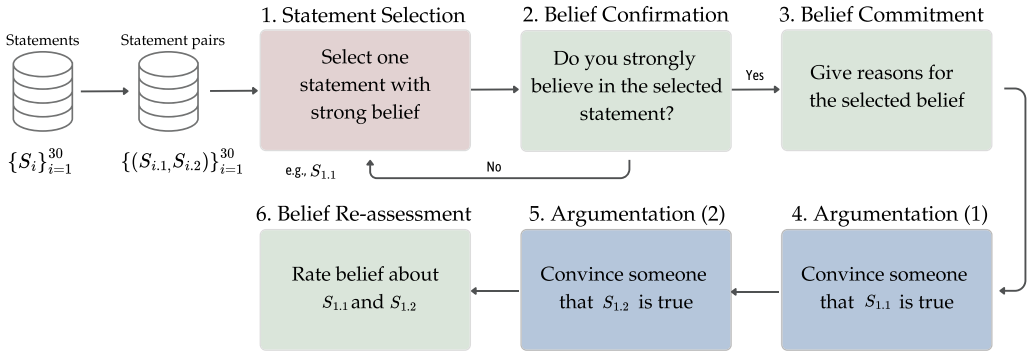


Figure 4
Survey design for *Accounting for Belief Change* condition (German corpus DEFABEL_V2_DE and English corpus DEFABEL_V2_EN).

4.3 Corpus III: Cross-linguistic Extension (English)

To examine whether the patterns observed in our belief-based deception framework generalize across languages, we extend our revised data collection protocol to English. This version mirrors the design used in the updated German corpus, including belief elicitation before and after argumentation, the use of bidirectional stimuli, and the incorporation of task elements intended to increase the personal relevance of the argumentation.

By aligning task design across both languages, we enable direct comparison of deceptive and truthful argumentation strategies in German and English. This setup allows us to explore potential language-specific variation in how deception is expressed, as well as the extent to which linguistic cues to deception are shared or divergent across typologically distinct languages.

The English corpus thereby serves both as a replication of our methodological improvements and as a testbed for cross-linguistic analysis within the belief-based deception framework.

4.4 Crowd-sourcing Study

In this section, we provide an overview of the data collection procedures for three corpora used in our study. While all corpora are based on user-generated content collected via online surveys, they differ in terms of language, recruitment platforms, and technical implementation.

4.4.1 ***Corpus I (DeFaBel_V1_De)****.* Participants are recruited via the online platform `Prolific`.[5] The study is implemented using `Google Forms`,[6] with all materials presented in German. Eligibility is restricted to individuals residing in Germany who report German as their native language, are fluent in German, and have an approval rate between 80% and 100% on the platform. There are no restrictions based on age or gender.

To ensure privacy, participants are not asked to provide any identifying information within the survey itself, and the collection of email addresses is explicitly disabled. Demographic information is only available to the extent that participants have consented to share it through Prolific. All participants receive compensation based on an estimated hourly rate of €10. If actual completion times differ from the estimate, payment is adjusted in line with Prolific's fair pay policy. In the argument generation study, participants are promised a bonus for producing convincing arguments; however, the bonus is granted to all participants regardless of argument quality. Reported payment includes this bonus.

Both belief assessment and argument generation studies include attention checks to verify that participants are engaged with the task. In the belief distribution study, 5 out of the total 55 items are explicitly marked as attention checks, requiring participants to select a specific value on the rating scale. In the argument generation task, participants are instructed to enter a predefined word instead of an argumentative text as a means of confirming task compliance.

5 `https://www.prolific.co/`.
6 `https://docs.google.com/forms/`.

The belief distribution study was conducted in September 2023, with 161 unique participants (including those from a pilot study). Participants are paid €1.74 for completing the full set of 55 questions. The argument generation task took place in October 2023 and included 171 participants. After excluding participants who did not consent to share demographic data, the average age of participants is 31.8 years (ranging from 19 to 72 years). Among those who provided gender information, 153 identify as male and 96 as female. On average, contributors received €0.90 per argument and an additional €0.58 bonus per submission. The total cost of the study amounted to approximately €1.62k.

4.4.2 ***Corpus II (DEFABEL_V2_DE).*** Participants are recruited via two online platforms, `Prolific` and `Clickworker`,[7] to ensure a sufficient sample size, as the number of texts written per participant is reduced and individuals with prior participation are excluded. The study is implemented using `Streamlit`,[8] an open-source Python framework, and hosted on in-house servers.[9] All materials are presented in German. Eligibility is restricted to native and fluent speakers of German residing in Germany. Each participant is allowed to complete the survey only once.

Due to a limited participant pool on Prolific, recruitment was extended to Clickworker. To prevent duplicate participation across platforms, a screening question is added asking participants to confirm that they had not previously taken part in a similar study. In total, 526 individuals completed the study. Each participant is asked to produce one deceptive and one truthful text. Demographic information for Prolific participants is drawn directly from platform data; for Clickworker participants, this information is collected within the survey. Disclosure of demographic data is optional. After excluding individuals who chose not to share demographic details, the average age of participants is approximately 34 years. Among those who provided gender information, 58% identify as male and 42% as female; a small number of participants opted not to disclose their gender.

The study was conducted between September 2024 and January 2025. Participants receive an effective hourly rate of approximately €9, plus a fixed bonus of €0.50. Although the bonus is advertised as a reward for producing convincing arguments, it is granted to all participants regardless of response quality. The total expenditure for the study amounts to approximately €1.6k.

4.4.3 ***Corpus III (DEFABEL_V2_EN).*** Participants are recruited via the online platform `Prolific`. The study uses the same survey design as in DEFABEL_V2_DE, with all materials and stimuli translated into English. Eligibility is restricted to fluent English speakers residing in the United Kingdom.

A total of 422 participants completed the study. As in the German version, each participant produced one deceptive and one truthful text. Among those who provided demographic information, 53% identify as male and 45% as female, while 0.47% preferred not to disclose gender information. For the remaining participants, gender data is unavailable. The average age of participants is approximately 39 years.

7 `https://www.clickworker.com`.
8 `https://streamlit.io`.
9 Implementation available at `https://bitbucket.org/aswathyve/argumentation-survey`.

The study was conducted between January and February 2025. Compensation is aligned with the German version of the study, with an effective hourly rate of approximately €9 and a fixed bonus of €0.50. The total cost of the study is approximately €1.2k.

4.5 Corpus Details

While reviewing the collected data, we identified a number of invalid responses, including instances where participants submitted irrelevant or nonsensical text. We also identified and removed instances where participants attended the survey more than once, either through self-disclosure within the survey or by comparing their unique identification numbers within platforms. Furthermore, some responses appeared to involve the use of text generation tools, which are flagged based on the time taken to write responses and formatting patterns typical of such tools, such as excessive markdown usage. For these questionable cases, participants are asked to return to the study, giving them an opportunity to refute our claim that they had used text generation tools. While this process helped identify many instances, it is still possible that some participants used such tools in ways that could not be reliably detected.

After applying quality control measures to exclude problematic instances, the final dataset comprises a total of 1,031 instances in DEFABEL_V1_DE, 968 instances in DEFABEL_V2_DE, and 804 instances in DEFABEL_V2_EN. Both DEFABEL_V2_DE and DEFABEL_V2_EN are *balanced* datasets, each containing an equal number of truthful and deceptive texts. Specifically, DEFABEL_V2_DE includes 484 truthful and 484 deceptive texts in German, while DEFABEL_V2_EN contains 402 truthful and 402 deceptive texts in English. This balanced distribution facilitates controlled cross-linguistic comparisons and helps mitigate potential biases in subsequent analyses. In contrast, DEFABEL_V1_DE displays a *skewed* label distribution, with 388 instances labeled as truthful and 643 as deceptive, resulting in approximately 62% deceptive texts.

With regard to text characteristics, DEFABEL_V1_DE exhibits an average of 4.75 sentences per instance and an average token count of 88.56, with lengths ranging from 16 to 262 tokens. Instances in DEFABEL_V2_DE are somewhat shorter, with an average of 3.87 sentences and 71.71 tokens per instance, ranging from 18 to 530 tokens. The English texts in DEFABEL_V2_EN show a similar sentence structure to DEFABEL_V2_DE, with an average of 3.77 sentences per instance, but a higher average token count of 89.80, ranging from 24 to 341 tokens.

Table 3 presents the descriptive statistics of all three DEFABEL datasets.

Table 3
Descriptive statistics for DEFABEL datasets.

Language	Dataset	Total	Truth	Decep.	**Avg. Sent.** (Min-Max)	**Avg. Token** (Min-Max)
German	DEFABEL_V1_DE	1031	388	643	4.75 (1–16)	88.56 (16–262)
	DEFABEL_V2_DE	968	484	484	3.87 (1–20)	71.71 (18–530)
English	DEFABEL_V2_EN	804	402	402	3.77 (1–18)	89.80 (24–341)
	CROSS_CULT_DE	1200	600	600	5.17 (1–16)	82.72 (7–302)
	OPEN_DOMAIN	7168	3584	3584	1.01 (1–5)	9.24 (2–60)
	DEC_OP	2500	1250	1250	4.03 (1–17)	65.03 (6–368)
	INTENTION	1640	783	857	2.70 (1–14)	54.84 (15–297)

5. Experiments

Our aim is to assess the reliability of linguistic cues commonly associated with deception and to evaluate the performance of a range of classification models, from feature-based approaches to large-scale language models. To understand the impact of controlling for belief change within the belief-based deception framework, we compare the two German DEFABEL datasets. To evaluate how belief-based deception differs from other forms of deceptive communication, we analyze the English DEFABEL dataset alongside a selection of established deception corpora. By contrasting patterns across datasets, we examine the extent to which deceptive intent can be identified through language alone. The section first outlines the datasets included in the comparison, followed by the evaluation experiments.

5.1 Datasets for Comparative Analysis

We use the three DEFABEL datasets, annotated using a belief-based deception framework, to investigate different aspects of deceptive communication. The German datasets, DEFABEL_V1_DE and DEFABEL_V2_DE, allow us to examine the effect of belief change control. The availability of both German and English DEFABEL datasets allows us to compare linguistic patterns across languages under belief-based deception framework. Additionally, the English DEFABEL dataset enables direct comparison between belief-based deception and alternative annotation schemes, given that the majority of existing deception datasets are available in English. For this comparison, we draw on four established corpora: CROSS_CULT_DE (Pérez-Rosas and Mihalcea 2014), DEC_OP (Capuozzo et al. 2020b), OPEN_DOMAIN (Pérez-Rosas and Mihalcea 2015), and INTENTION (Kleinberg and Verschuere 2021).

The CROSS_CULT_DE and DEC_OP datasets are opinion-based deception datasets, where participants are asked to argue *for* or *against* their own stance on controversial, morally charged issues such as abortion, euthanasia, or cannabis legalization. CROSS_CULT_DE additionally includes a more affective prompt about friendships. These datasets are similar to DEFABEL in that they are also argumentative in nature; however, they do not require participants to be persuasive and are grounded in personal opinions rather than beliefs. The OPEN_DOMAIN dataset comprises open-ended factual statements, where participants wrote seven truthful and seven deceptive sentences on any topic of their choice, provided the content remained plausible, reflecting everyday lying behavior. Finally, the INTENTION dataset consists of autobiographical narratives about the most important non-work-related activity participants planned for the upcoming week. Depending on the assigned condition, participants described this activity truthfully or deceptively. This dataset captures the role of personal involvement and narrative structure in deceptive expression.

Table 3 summarizes descriptive statistics for all datasets. To assess their suitability for comparative analysis, we compute pairwise corpus similarity using the method proposed by Li and Dunn (2022), which compares unigram and character trigram distributions via Spearman's ρ.[10] As shown in Figure 5, the opinion-based datasets DEC_OP and CROSS_CULT_DE show the highest similarity ($\rho = 0.83$). DEFABEL_V2_EN exhibits moderate similarity with the other datasets ($\rho = 0.58$–0.68), supporting its inclusion in comparative experiments. Overall, these values suggest that the datasets

10 Implementation available at: `https://github.com/jonathandunn/corpus_similarity`.

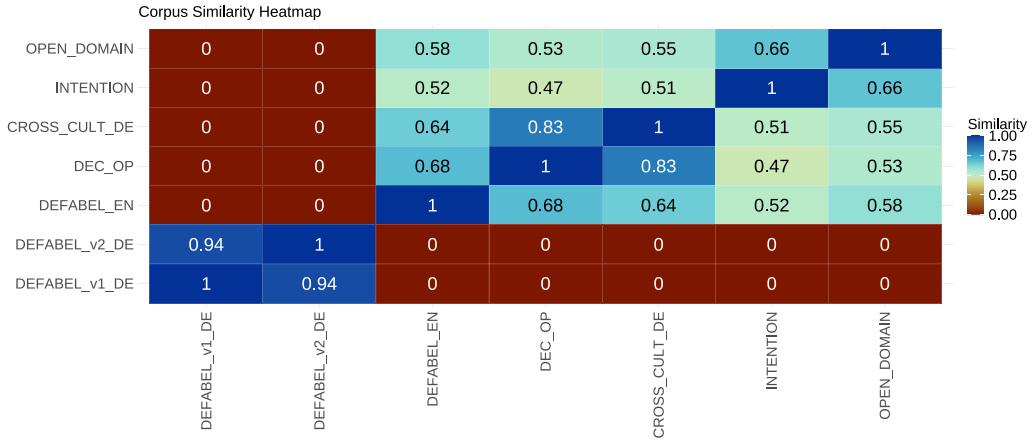


Figure 5
Pairwise similarity values, based on Li and Dunn (2022).

share significant overlap in term and character distributions, thereby validating their suitability for comparative analysis.

By comparing these datasets, we aim to gain a better understanding of the factors influencing deception detection performance and determine whether belief-based deception exhibits fundamentally weaker deception cues than other forms of deceptive text.

All datasets considered here represent relatively low-stakes forms of deception, where participants have no personal or material consequences for being deceptive. This focus reflects both the nature of available corpora and the types of contexts most relevant to NLP applications, such as online discussions, reviews, or everyday persuasive writing. High-stakes deception, such as in legal, forensic, or financial documents, differs fundamentally in motivation and consequence, making it unsuitable for direct comparison and rarely accessible for research. Our analysis therefore focuses on forms of written deception that is representative of how deception typically appears in everyday textual communication.

5.2 Are Linguistic Cues Reliable Indicators of Deception?

To examine the reliability of linguistic cues commonly associated with deception and to establish a foundation for assessing their predictive utility, we conduct a correlation analysis. We aim to quantify the strength and significance of relationships between individual linguistic cues and deception labels.

5.2.1 Experimental Setup. We identify linguistic indicators of deception based on prior literature. Guided by the linguistic categories outlined in Table 1, individual features representing each category are derived using the LIWC lexicon (e.g., affective expressions, cognitive processes, self-references), readability measures (e.g., Flesch-Kincaid and Gunning-Fog indices), and other statistical metrics, including word count, lexical diversity, and part-of-speech frequency. These linguistic cues are systematically computed for each text using validated linguistic feature extraction tools (like LIWC) and standard Python libraries. A detailed description of the complete feature set is available in Appendix Table B.1.

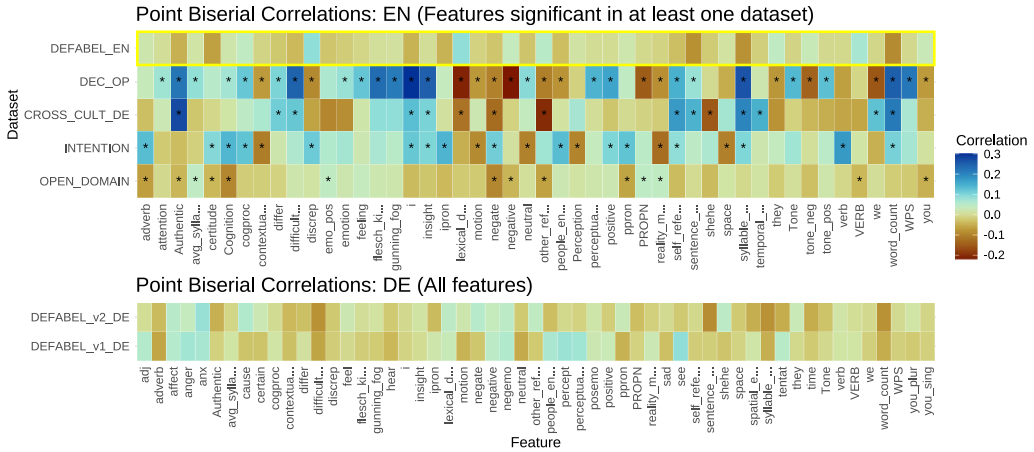


Figure 6
Point biserial correlations between linguistic features and deception labels across English (top) and German (bottom) datasets.

For LIWC lexicon-based measures, we used LIWC-22 for English, and the DE-LIWC2015 for texts in German. The feature labels differ slightly between the two lexicons, with the German lexicon containing fewer features compared with the English version. Additional linguistic features are computed using Python-based text-analysis libraries, including `spaCy`, `germansentiment`, `vaderSentiment`, `textstat`, `NLTK`, and `pyphen`. Code and resources for generating these linguistic features are openly accessible via our Bitbucket repository.[11]

Correlation Analysis. To identify linguistic features correlated with deception, we compute point-biserial correlations (Glass and Hopkins 1996) between each linguistic feature and the binary deception labels for each dataset. The features in all datasets are standardized using z-score normalization ($\mu = 0$, $\sigma = 1$) to ensure comparability across datasets. Features that lack sufficient variance or did not meet statistical assumptions (e.g., constant predictors or insufficient observations) are excluded from correlation computations. To control for multiple hypothesis testing, Bonferroni correction (Bonferroni 1936) is applied separately within each dataset. Correlations with an adjusted p-value below 0.05 are considered statistically significant.

We conduct the analysis using the statistical programming language R (R Core Team 2024), using packages `dplyr` for data manipulation, `purrr` for functional programming, and `stringr` for efficient text processing.

5.2.2 Results. The results of the correlation analysis are visualized as heatmaps in Figure 6, showing the magnitudes and directions of correlations across datasets and linguistic features. Statistically significant correlations are marked with an asterisk (*).

A clear pattern can be observed on comparing datasets in the plot: the DeFaBel datasets in both English and German display predominantly weak correlations, with none of the features showing any statistical significance. This finding aligns with our prior observations reported in Velutharambath, Wührl, and Klinger (2024b), where similar analysis on the DeFaBel_V1_De dataset also shows an absence of significant

11 `https://bitbucket.org/aswathyve/linguistic_style_features/`.

linguistic cues. Across all datasets, correlation coefficients range approximately between 0.3 and −0.2, indicating relatively weak overall relationships. Nevertheless, other English datasets exhibit weak yet statistically significant correlations for some features.

The DEC_OP dataset exhibits comparatively stronger correlations consistent with prior research. Specifically, categories such as authenticity (*authentic*), self-reference (*I*, *we*), and text quantity (*word_count*, *WPS*, *syllable_count*) correlate positively with truthfulness, while negative sentiment (*negative*) and references to others (*other_reference*, *they*) are indicative of deception. Similar trends emerge in the CROSS_CULT_DE dataset, likely reflecting its methodological similarity to the DEC_OP dataset.

Contrary to the expectations based on prior research, *lexical diversity* correlates positively with deception rather than truthfulness in most datasets (DEC_OP, CROSS_CULT_DE, INTENTION), although notably not in the DEFABEL datasets. While the authenticity feature (*authentic*), which is usually an indicator of truthfulness, does so in opinion-based datasets (DEC_OP, CROSS_CULT_DE), it correlates with deception in open domain and personal narratives (OPEN_DOMAIN, INTENTION, resp.) datasets. The OPEN_DOMAIN dataset shows generally weak correlations, comparable to the DEFABEL datasets, yet some correlations still reach statistical significance.

Overall, the correlation analysis questions the reliability of linguistic indicators of deception and emphasizes the strong influence of data collection methodology on the strength of linguistic cues. Datasets collected using similar protocols (DEC_OP and CROSS_CULT_DE) exhibited comparable correlation patterns. In contrast, the absence of statistically significant linguistic indicators within the DEFABEL datasets—despite efforts to enhance ecological validity through a revised data collection procedure—suggests that weaker cues result from the belief-based deception framework rather than language-specific factors.

5.3 How Accurately Can Models Detect Deception?

To systematically evaluate the predictive value of linguistic features in deception detection, we conduct experiments using (1) traditional machine learning models trained on linguistic features, (2) Transformer-based language models fine-tuned for classification, and (3) instruction-tuned large language models applied in a zero-shot setting. This setup allows for a comparison between models that rely on predefined linguistic cues, those that learn representations from raw text, and models that follow task instructions without additional supervision.

5.3.1 Experimental Setup.

Datasets and Splits. Our experiments are conducted on the selected set of deception datasets discussed in Section 5.1, which includes DEFABEL_V1_DE and DEFABEL_V2_DE in German, and DEFABEL_V2_EN, CROSS_CULT_DE, DEC_OP, OPEN_DOMAIN, INTENTION in English. For each of the datasets, we follow a consistent topic-disjoint data partitioning strategy, splitting the data into a training set for model development and a holdout set reserved for final evaluation. This ensures that no overlapping topics appear in both subsets, thereby evaluating models' generalization to unseen instances. See Appendix Table B.4 for distribution of labels across the different splits of data.

Linguistic Feature-based Classification. To assess the predictive value of linguistic cues as features, we train feature-based classifiers for the task of deception detection. As

features we included all of them used for the correlation analysis. They span a range of theoretically motivated dimensions, including syntactic patterns, affective and cognitive expressions, lexical diversity, readability, and referential strategies (see Appendix Table B.1).

We implement four feature-based classifiers—Logistic Regression, Random Forest, Support Vector Classifier (SVC), and Gradient Boosting—chosen for their diversity in learning approaches, from simple linear models to complex ensemble methods and non-linear classifiers. This variety ensures reliable evaluation across different data patterns. All features undergo z-score normalization to ensure comparability across dimensions. To evaluate model robustness and reduce sensitivity to data splits, we conduct stratified 10-fold cross-validation, preserving class distributions. After cross-validation, each classifier is retrained on the entire training set and evaluated on the topic-disjoint holdout set to assess generalization beyond memorized content. We use the `scikit-learn` python library for this implementation.

Transformer-based Text Classification. To evaluate the effectiveness of Transformer-based models in the deception detection task, we fine-tune models from the BERT (Devlin et al. 2019) family. For English datasets, we use `FacebookAI/roberta-large`, trained on a broad English corpus, and for German datasets, we utilize `TUM/GottBERT_base_best`, pretrained on high-quality German texts.

Transformer models are fine-tuned by adding a classification layer and optimizing cross-entropy loss using the AdamW optimizer with a fixed learning rate of $1e{-}5$. Early stopping based on validation loss with a patience of five epochs mitigates overfitting, a critical step given the limited size and domain-specific nature of our datasets. Training is conducted for up to eight epochs with a batch size of 8, reflecting computational constraints and available data.

Model performance is evaluated through 10-fold cross-validation on the development set. All Transformer-based experiments are implemented using the `transformers` library.

Baselines. In addition to the feature-based classifiers, we include two standard baselines to establish meaningful lower bounds for model performance. The majority class baseline is designed to always predict the most frequent label in the training data. While most datasets in our benchmark are balanced, DEFABEL_V1_DE exhibits a notable skew toward the deceptive class, so as a baseline we predict all instances to be "deceptive". The random baseline, by contrast, assigns labels uniformly at random. These baselines serve as critical reference points, enabling us to assess whether models are learning informative patterns or merely capitalizing on label distribution.

Prompting with LLMs. Given the substantial computational and financial costs of running LLM inference on a large dataset (≈15k texts) across multiple prompt configurations and models, we use a multi-stage evaluation strategy. In the first stage, we assess model performance using a simple zero-shot prompt with smaller models, evaluating results on both the training and holdout sets to ensure comparability with traditional classifiers. To examine the effect of prompt variation, we then test a set of prompt variants, this time restricted to the holdout dataset, to capture differences in instruction phrasing without excessive computation. This approach leverages the fact that instruction-tuned LLMs can be evaluated directly without additional task-specific training. Finally, to investigate the influence of model size, we conduct additional experiments on the holdout dataset using models of varying scales.

(1) **Simple Prompt Set-up:** We use three open-source instruction-tuned models: `Mistral-7B-Instruct`,[12] `Llama-3.1-8B`,[13] and `Qwen1.5-14B-Chat`.[14] `Mistral-7B-Instruct` is a lightweight model known for strong instruction-following capabilities and competitive multilingual performance, including German (Jiang et al. 2024). `Llama-3.1-8B` complements Mistral in size but differs in model design and training setup, and has been reported to perform well in generalization and reasoning, particularly in zero-shot tasks (Grattafiori et al. 2024). `Qwen2.5-14B-Instruct`, by contrast, is a larger model with enhanced multilingual coverage and improved reasoning abilities, particularly in complex language understanding tasks (Bai et al. 2023). Additionally, we include `GPT-4o-mini`,[15] a proprietary model, to provide a comparative perspective against open-source alternatives.

At this stage, our aim is not to identify the most effective or best-performing prompt for deception detection, but rather to explore how instruction-tuned language models interpret deception-related tasks when prompted in a zero-shot setting. We use the following zero-shot prompt for evaluating all models:

> **System:** *You are an expert AI assistant who can detect truthful and deceptive intention in text very accurately.*
>
> **User:** *Evaluate whether the author of the provided text is being deceptive. Consider the author to be deceptive if they intentionally state something they know to be untrue or deliberately misrepresent their thoughts, feelings, intentions, or actions. Take into account that the author could hold false beliefs and hence intention is more important than factual accuracy.*
> *Provide your judgment strictly in English in the following JSON format:*
> *{ "is_lying": true/false, "reason_title": Very brief summary (as a short phrase) of why you believe the author is or isn't lying, "explanation": Detailed explanation supporting your judgment }*
> *Text to evaluate:* `<input_instance>`

In the prompt, we explicitly instruct the models to assess whether the author of a given text is lying. The instruction defines deception as the intentional misrepresentation of truth. It also clarifies that the author may hold false beliefs, and hence, the model is asked to prioritize the author's intent over factual accuracy. Additionally, each model is directed to produce (1) a brief phrase summarizing the reason for its decision (the *reason title*) and (2) a detailed explanation outlining its rationale. The goal here is to observe LLM behavior under a simplified prompt setting without any prompt optimization efforts.

12 https://huggingface.co/mistralai/mistral-Instruct-v0.3.
13 https://huggingface.co/meta-llama/Llama-3.1-8B.
14 https://huggingface.co/Qwen/Qwen2.5-14B-Instruct.
15 https://platform.openai.com/docs/models/GPT-4o-mini.

(2) **Prompt Variations:** We design a set of prompt variants to explore how different prompting strategies influence model behavior and reasoning patterns in deception detection:

- **Zero-shot:** This variant provides more detailed guidance on the reasoning process, instructing the model to assess deception by considering factors such as internal contradictions, strategic framing, emotional consistency, and contextual plausibility.
- **Random few-shot:** In this setting, two truthful and two deceptive examples are randomly selected and included to illustrate each class, offering the model minimal exposure to task-relevant examples before inference.
- **Feature-similarity few-shot:** Here, examples are chosen based on linguistic feature similarity to the input text. This approach allows the model to reference samples with comparable stylistic and structural characteristics, potentially improving contextual alignment.
- **Semantic-similarity few-shot:** In this variant, examples are selected using cosine similarity between sentence embeddings derived from a multilingual sentence transformer model.[16] This method ensures that the model observes examples that are semantically closest to the target text.
- **Chain-of-thought (CoT) reasoning:** Each of the above prompts is further extended with a CoT version, where the model is explicitly instructed to reason through the text before producing its final judgment. The CoT prompts guide the model to first reflect on the content, then articulate its reasoning, and finally provide the deception label.

In total, eight prompt configurations are evaluated in this stage (See Appendix section B.3). To maintain computational efficiency, we select the two best-performing open-source models from the initial evaluation, along with the proprietary model, for the prompt variation experiments.

(3) Model Size Variations: To further investigate the impact of model scale while maintaining computational efficiency, we evaluate the best-performing prompt on larger variants of the top-performing open-source and proprietary models. Specifically, we consider `Mistral-8×7B-Instruct`,[17] `Llama3.3-70B`,[18] and `GPT-4o`[19] as the larger counterparts of their respective architectures. This analysis aims to determine whether the models' deception detection capabilities scale with increasing parameter counts.

5.3.2 Results.

(a) Comparison Across Classifiers. Figure 7 shows the trade-off between F_1 scores for the deceptive and truthful classes (top), evaluated via 10-fold cross-validation across

16 `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`.
17 `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`.
18 `https://console.groq.com/docs/model/llama-3.3-70b-versatile`.
19 `https://platform.openai.com/docs/models/gpt-4o`.

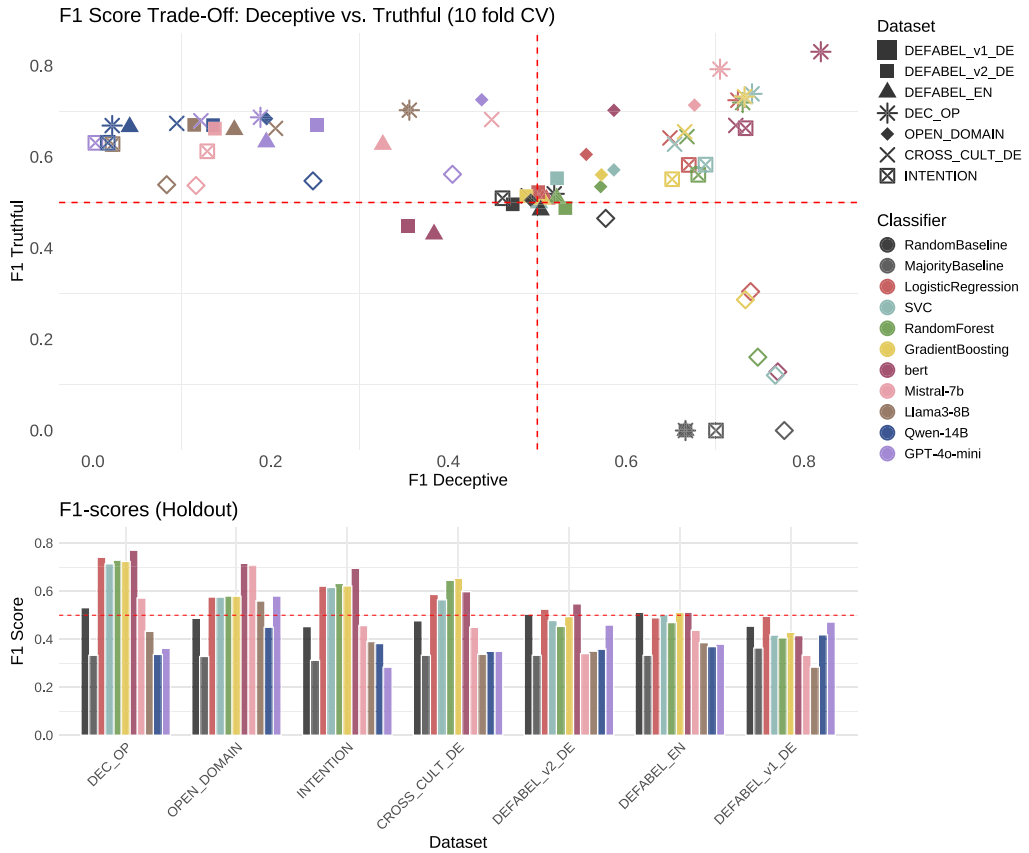


Figure 7
F_1 scores across datasets and models. Top: Trade-off between *deceptive* and *truthful* F_1 scores (10-fold cross-validation on development set). Bottom: Overall macro-F_1 scores on holdout sets. The dotted red lines are included for visual reference, indicating that the macro-F_1 scores for the random baselines are centered around 0.50.

different datasets and models. The LLM results correspond to the *simple prompt set-up*. The bottom panel of the figure shows the corresponding model performance on the holdout sets, which consist of unseen instances. These holdout sets are used strictly for evaluation purposes and not for model optimization; all model parameters are fixed prior to this evaluation. Unless stated otherwise, all reported scores are macro-averaged F_1 scores. See model performances across all datasets in the Appendix: Tables B.5, B.6, B.7, and B.8.

Detection of Belief-based Deception is Challenging. In Figure 7, we observe that among all the datasets, DEFABEL_V1_DE (♢) proves to be the most difficult for the models to classify accurately. Most models tend to predict a larger proportion of instances as deceptive, resulting in relatively high F_1 scores for the deceptive class. This pattern closely resembles the majority class baseline, which predicts all instances as deceptive. Especially for the fine-tuned BERT model, this behavior is consistent across folds, suggesting a systematic modeling issue rather than random variation. On the holdout set, model performance on DEFABEL_V1_DE drops substantially, with macro F_1 scores

falling below those of the random baseline for all models. We further observe that both DEFABEL_V2_DE (■) and DEFABEL_V2_EN (▲) achieve model performance close to the random baseline, with the fine-tuned BERT model underperforming compared to the feature-based classifiers. In comparison to the DEFABEL_V1_DE, the BERT model performs better on labeling truthful instances. This observation suggests that accounting for belief change during argumentation could have improved the data quality or validity. However, detecting deception embedded within otherwise truthful content remains particularly difficult even for finetuned models with more contextual information. One plausible explanation for BERT's behavior is its pretraining on large-scale corpora, which may bias the model toward associating plausible language with truthful intent—even when communicative intention is deceptive.

Within-dataset Regularities Result in Better Performance. All English datasets, except DEFABEL_V2_EN, exhibit a similar pattern: Feature-based models achieve moderate to strong classification performance, and fine-tuning BERT consistently improves results. This suggests that models are able to leverage consistent linguistic patterns in these datasets, enabling more reliable differentiation between deceptive and truthful instances. The performance of models on DEC_OP dataset is particularly noteworthy. On both development and holdout data, all models (except LLMs) achieve relatively high F_1 scores (0.7–0.8), with BERT achieving close to 0.8 for both the deceptive and truthful classes. While BERT outperforms the feature-based models on this dataset, the feature-based models themselves still achieve strong results, markedly better than their performance on the other datasets. This consistent success across model types is especially notable given the inherently subtle nature of deception cues in textual context. The results suggest that the DEC_OP dataset may contain linguistic or stylistic patterns that align closely with class labels—patterns that are more easily detectable than those in other corpora. This interpretation is supported by our linguistic feature correlation analysis, which revealed subtle yet statistically significant associations between deception labels and multiple linguistic cues. Together, these findings suggest that the observed separability in DEC_OP may result more from annotation consistency or structural regularities than from generalizable deception indicators.

The results raise important questions about what the models are truly learning to detect. While performance appears strong on certain datasets, we assume that models may be exploiting dataset-specific regularities, such as stylistic or structural patterns, rather than genuinely identifying deceptive intent. A consistent pattern across our analysis is that deception becomes much harder to detect when it is embedded within factually accurate context. This is particularly evident in the more challenging DEFABEL datasets, where deceptive intent does not manifest through clear or consistent linguistic markers. Instead, it is often subtle, context-dependent, and inseparable from truthful language. This exposes a core limitation of current models: their success appears to depend on the presence of stable surface-level patterns that coincide with deception labels. When such patterns are absent or weak, the models fail to distinguish deceptive intent, raising concerns about what these systems are actually learning to recognize.

LLMs Show a Pronounced Truth Bias with Simple Prompts. When analyzing the performance of the LLM-based classifiers in the *simple prompt setting*, we observe that all models—Mistral-7B, LLaMA-8B, Qwen-14B, and GPT-4o-mini—generally perform worse than both the feature-based and fine-tuned Transformer models across most datasets. An exception is Mistral-7B, which performs comparatively well on the DEC_OP and OPEN_DOMAIN datasets. However, in the DEFABEL datasets, all

instruction-tuned models exhibit a consistent and pronounced tendency to label instances as truthful, regardless of the actual author intent. This supports our hypothesis that instruction-tuned LLMs, optimized for alignment and helpfulness, are not well-suited to deception detection tasks. Rather than critically evaluating authorial intent, these models may rely heavily on factual plausibility, leading to a systematic bias toward truthful classifications. This tendency likely limits their reliability in detecting deceptive intent and is clearly reflected in the rationales provided by the models for their predictions (see Section 5.4).

(b) Comparison Across Prompt Variants. Among the open-source LLMs, Qwen-14B consistently exhibits the weakest performance across datasets. Therefore, the prompt variation experiment includes LLaMA-8B and Mistral-7B as the representative open-source models, alongside GPT-4o-mini as the proprietary model. Figure 8 presents a comparison of F_1 scores—deceptive, truthful, and overall—across different prompt variants for the three models and seven datasets. See Tables B.9, B.10, and B.11 for F_1 scores, precision, and recall values across prompt variants for each dataset.

LLMs Exhibit Variable Bias Patterns Across Prompt Variants. When comparing the simple zero-shot prompt (*zs_simple*) with the more detailed version (*zs_detailed*), we observe that

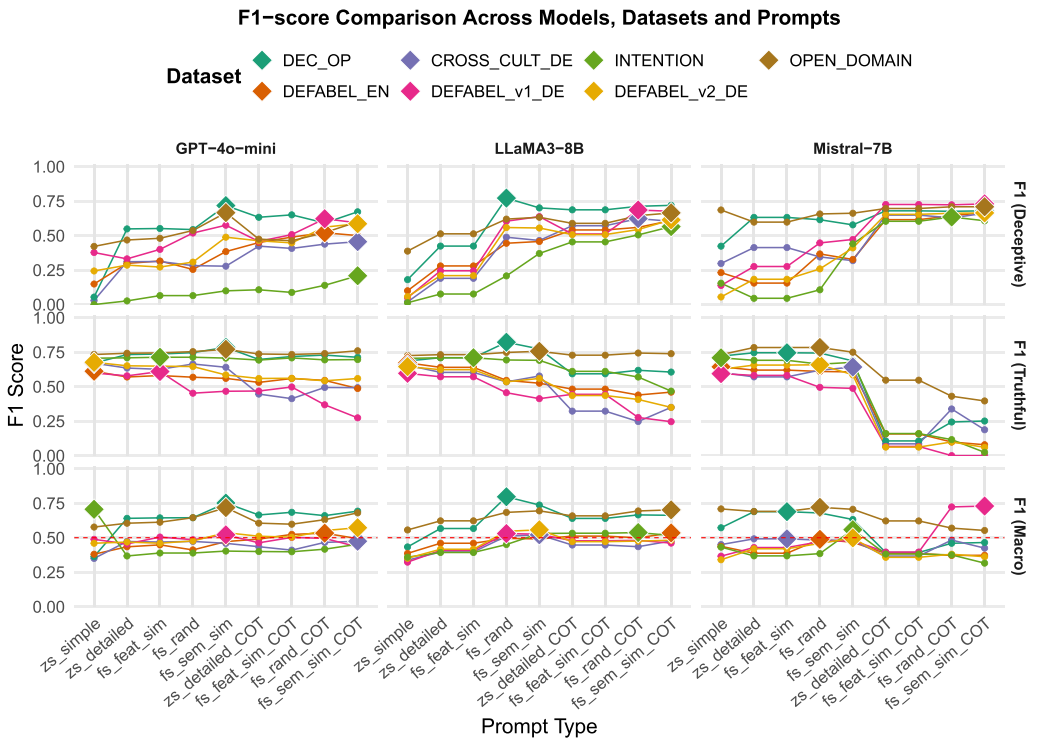


Figure 8
F_1-score comparison across prompt variants for models: GPT-4o-mini, LLaMA-8B, and Mistral-7B, evaluated on seven datasets. Circular markers indicate individual prompt variants, while diamond markers highlight the best-performing prompt per model–dataset–metric combination. The red dashed line denotes the 0.5 macro-F_1 chance baseline.

the latter generally achieves higher overall F_1 scores across models and datasets, with only a few exceptions. For example, the best performance on INTENTION is observed with the simple prompt, and on the DEFABEL dataset performance either remains the same or worsens for GPT-4o-mini. Adding more examples to the detailed prompt does not substantially improve model performance overall. However, for DEC_OP and OPEN_DOMAIN, we find that incorporating semantically similar instances as few-shot examples (*fs_sem_sim*) leads to a notable boost in performance ($F_1 \geq .75$ for GPT-4o-mini and LLaMA-8B, and $\geq .7$ across all models), approaching the high scores reported for fine-tuned BERT models. This finding reinforces the hypothesis that model performance is driven in part by dataset-level regularities, given that deception cues are often embedded in stylistic rather than semantic features.

With the simple prompt, we observe that LLMs tend to exhibit a truth bias. However, this trend reverses when CoT reasoning is introduced into the prompts. A consistent pattern in Figure 8 shows that F_1 scores for the deceptive label increase, while those for the truthful label decrease under CoT-based prompts compared to non-CoT prompts. This suggests that CoT prompting induces a bias toward labeling more instances as deceptive—a pattern particularly pronounced in Mistral-7B compared with the other models.

These patterns are also evident in the DEFABEL datasets. However, the best F_1 scores in this case still range only between .53 and .57, with the exception of a CoT prompt ($F_1 = .73$), which is largely an artifact of predicting all instances as deceptive. Overall, these observations suggest that including more information in the prompts generally improves model performance, yet no single prompting strategy yields consistent improvements across models or datasets. CoT prompting tends to bias models toward predicting instances as deceptive. Furthermore, the overall F_1 scores indicate that LLMs still struggle to achieve performance substantially above a random baseline. This remains true for all datasets except for DEC_OP and OPEN_DOMAIN.

(c) Comparison Across Model Sizes. Building on the insights from the previous experiments, we next examine how model size influences deception detection performance. For this analysis, we focus on the LLaMA and GPT model families to assess whether larger architectures are better able to capture stylistic cues indicative of deception. We evaluate the prompts *zs_detailed*, *fs_random*, and *fs_sem_sim*, along with their CoT variants, as these configurations showed improved performance across datasets and models, albeit not consistently.

From Figure 9, we observe that increasing model size does not lead to a substantial improvement in overall F_1 scores. Interestingly, in larger models, the truth bias observed with the simple prompt re-emerges across all prompt types. The deceptive-label bias observed for CoT prompting in smaller models is largely absent. These results suggest that scaling up model size mitigates some of the prompt-specific biases observed in smaller architectures but does not yield notable gains in deception detection accuracy.

(d) Cross-linguistic Comparison on* DeFaBel *Datasets. To assess whether the observed bias patterns generalize across languages, we compare model performance on the English (DEFABEL_V2_EN) and German (DEFABEL_V1_DE, DEFABEL_V2_DE) variants of the DEFABEL datasets. Overall, the results show very similar performance trends across languages, with differences primarily in magnitude rather than direction. In both English and German, the bias patterns observed for prompts and model sizes also hold. While the general behavioral patterns are comparable across languages, the German datasets—especially DEFABEL_V1_DE—exhibit larger bias swings and greater

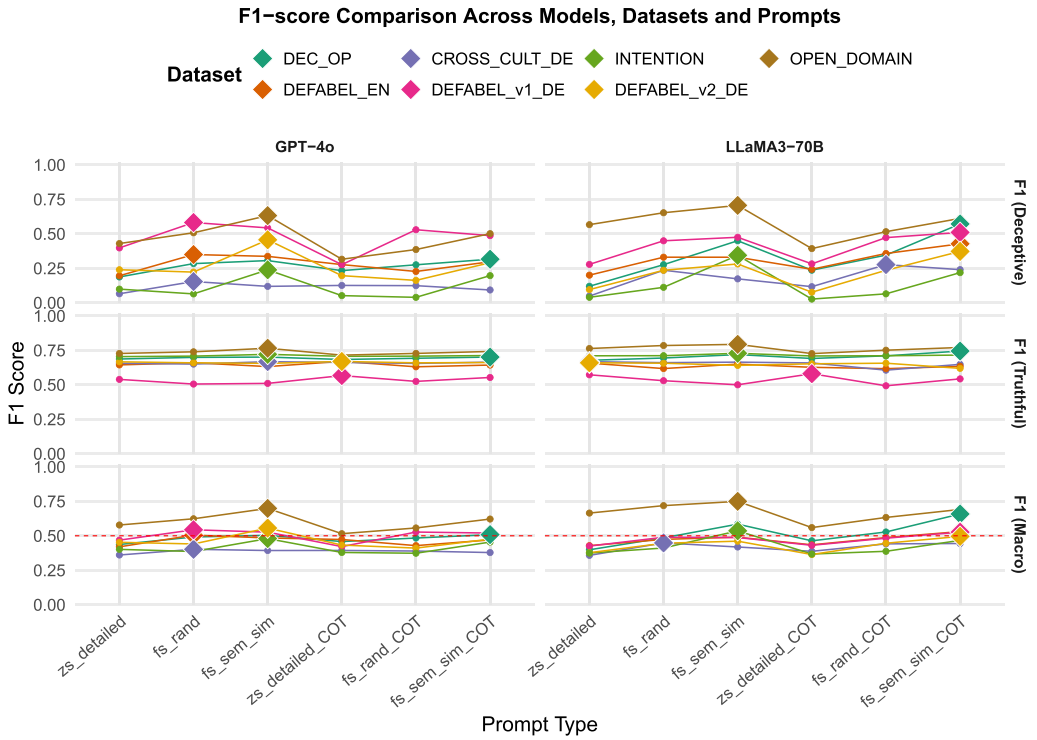


Figure 9
F_1-score comparison across prompt variants for larger language models: GPT-4o and LLaMA3-70B, evaluated on seven datasets. Circular markers indicate individual prompt variants, and diamond markers highlight the best-performing prompt per model–dataset–metric combination. The red dashed line denotes the 0.5 macro-F_1 chance baseline.

performance variability across prompt types. CoT prompts, in particular, often produce much higher false-positive rates in German compared with the English dataset. However, these differences are relatively minor and may stem from the predominance of English in model pretraining and instruction-tuning data, as well as from structural and stylistic differences in how deceptive cues are expressed in German. Based on these findings, we argue that the observations made on belief-based deception are largely language-independent.

5.4 How Do LLMs Justify Misclassifications?

Deception detection experiments with instruction-tuned models revealed that variations in prompts and model size do not lead to consistent improvements in overall performance. In the zero-shot simple prompt, the models were not provided with explicit cues about deception-detection strategies such as internal contradiction, strategic framing, emotional consistency, or contextual plausibility. This setup offers a baseline for examining how models justify their predictions.

In this section, we examine the internal reasoning patterns that underlie incorrect predictions made by large language models in the context of deception detection. The analysis focuses on the explanations that LLMs generate when they misclassify

inputs, specifically, when deceptive content is labeled as truthful or vice versa. By systematically analyzing these justifications, we aim to uncover recurring biases in model reasoning and assess whether they correspond to linguistic or pragmatic cues commonly associated with deceptive language.

5.4.1 Experimental Setup. As discussed in Section 5.3.1, our LLM-based classification experiments involve prompting instruction-tuned models to perform deception detection. Given an input text, each model is asked to output (1) a binary label indicating whether the author is lying or telling the truth, (2) a short justification summarizing the rationale for its decision (referred to as the *reason title*), and (3) a longer, more detailed explanation of the decision process.

Figure 10 shows the precision, recall, and macro F_1 for each LLM evaluated on the full dataset. On reviewing the results, we observe a consistent pattern across all LLMs: a strong bias toward predicting inputs as truthful. We see that recall for the truthful class is consistently high, indicating that the models correctly identify most truthful instances. However, this comes at the cost of lower precision for the same class, suggesting that many deceptive inputs are incorrectly labeled as truthful. In contrast, for the deceptive class, precision is relatively high—when the model does predict deception, it is often correct—but recall is substantially lower, meaning the majority of deceptive instances are not detected. This imbalance reveals a systematic tendency to favor the truthful label unless there are clear and strong cues suggesting deception. To better understand the models' internal reasoning processes and sources of error, we conduct a detailed analysis of the misclassified instances.

Although our experiments include multiple LLMs, we center this error analysis on the outputs of the Mistral model as it consistently achieves stronger performance across datasets, both in terms of macro-averaged F_1 score and recall for deceptive instances.

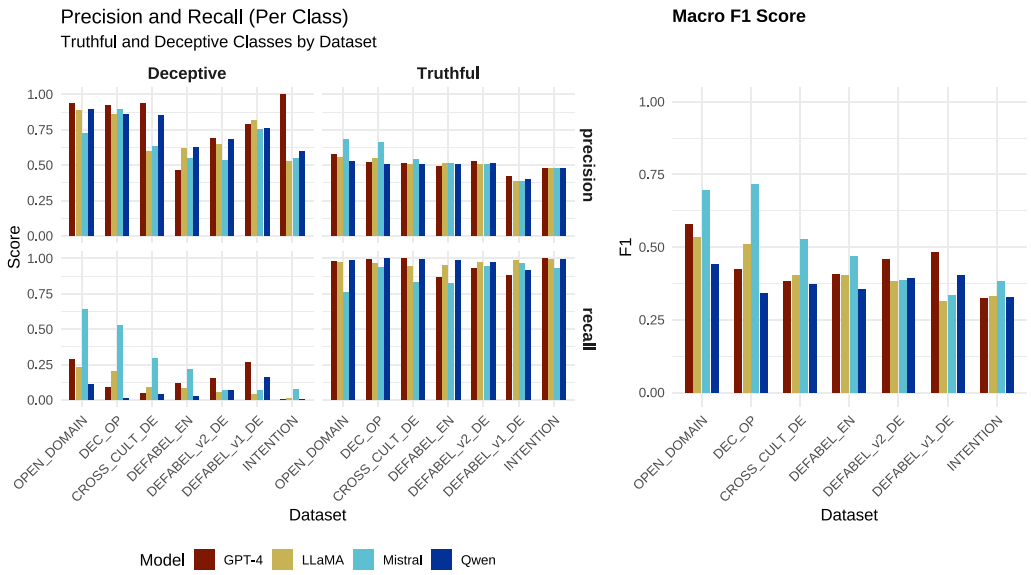


Figure 10
(Right) Precision vs. Recall comparison for Truthful and Deceptive labels across four LLMs, Mistral-7B, Llama-8B, Qwen-14B, and GPT-4o-mini (Left) macro F_1 scores for all LLMs.

These properties make it the most informative model for analyzing misclassification behavior, particularly in examining how deceptive content is interpreted and justified. We focus our analysis on the *reason titles*, treating them as concise, high-level representations of the models' decision rationales.

Each model prediction is compared to the corresponding ground truth label, and the outcomes are categorized into four standard evaluation classes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). In our evaluation, as the task is deception detection, we treat the *deceptive* class as the positive class. In this context, a false positive denotes a truthful input incorrectly classified as deceptive, while a false negative refers to a deceptive input wrongly labeled as truthful.

To investigate the reasoning patterns underlying model misclassifications, we focus on the *reason titles* associated with false positives and false negatives, as these errors offer the most direct insight into model failures. Each reason title is embedded using the `all-MiniLM-L6-v2` model from the SentenceTransformers[20] library, which maps the short justifications into dense semantic vectors in a high-dimensional space. We then apply KMeans clustering to group together semantically similar justifications within each error type. The number of clusters is determined dynamically for each subset, with an upper bound of 20 and a lower bound of 2, depending on the number of available samples. This allows us to maintain a balance between interpretability and resolution, while adapting to the size of each group. To visualize the resulting clusters, we reduce the embedding dimensions to two using t-SNE. The perplexity parameter for t-SNE is set adaptively based on subset size, capped at 30 and floored at 5, to ensure stability across datasets. The resulting two-dimensional coordinates are used to analyze and interpret the distribution of reasoning patterns across error types.

After clustering, we manually reviewed each group of reason titles within the false positive and false negative subsets. For each cluster, we examined representative examples and assigned a descriptive label summarizing its dominant reasoning pattern. In cases where multiple clusters reflected closely related themes, we merged them to avoid redundancy and improve interpretability. This qualitative analysis allowed us to identify recurring justification strategies used by the model when it misclassified deceptive or truthful instances.

5.4.2 Results. Figure 11 shows the proportion of reasoning clusters associated with false positives and false negatives, as produced by the Mistral model across datasets. Each bar represents the relative share of explanation types within a given error category, allowing comparison of dominant model rationales across different sources (see Table B.14 in the Appendix for absolute counts).

Among false negatives—deceptive instances wrongly predicted as truthful—the most common reasoning clusters include *Honest Tone with No Deception*, *Expressing Emotions*, and *Consistency and Coherence*. These features have been associated with truthful communication in prior work, where positive tone and emotion, and coherent narratives, are often correlated with honesty (Newman et al. 2003; Vrij 2008). As a result, deceptive messages that exhibit these characteristics are more likely to be misclassified as truthful by the model.

Consistent with our hypothesis that LLMs with large-scale pretraining are biased toward factual accuracy, we observe a strong reliance on factual plausibility as a cue for truthfulness. This tendency is particularly evident in the DEFABEL datasets and in

20 `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`.

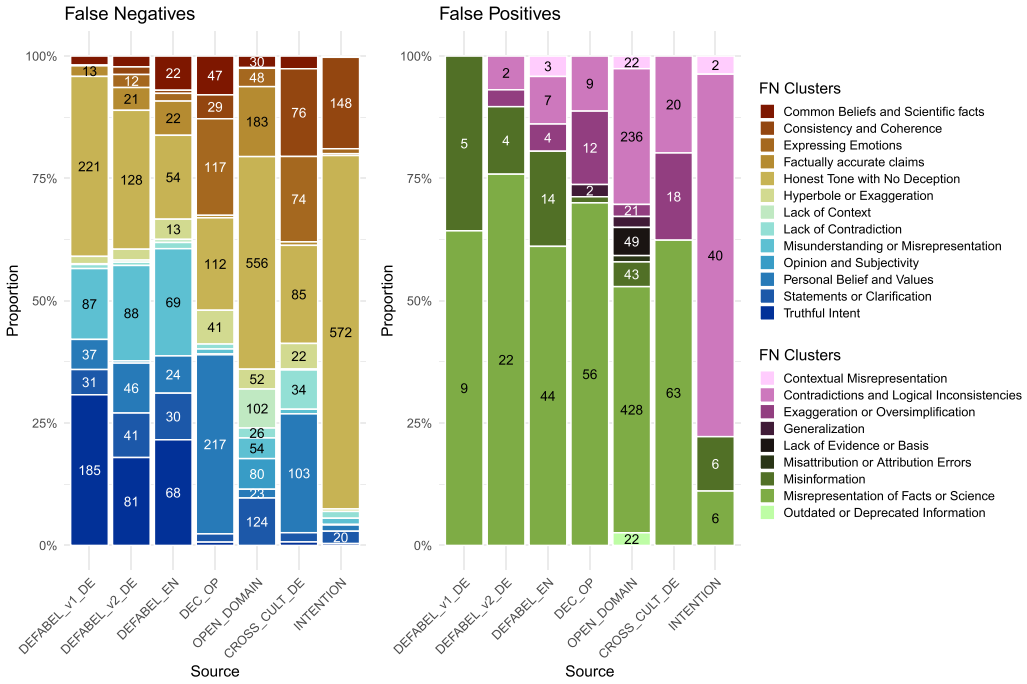


Figure 11
Distribution of explanation clusters among false positives (FPs) and false negatives (FNs) across datasets. Each bar shows the proportion of error instances attributed to different types of model reasoning.

OPEN_DOMAIN, where many deceptive instances are misclassified with justifications such as *Factually Accurate Claims* or *Misunderstandings or Misrepresentations*. These explanations indicate that the model focuses on factual correctness over the speaker's underlying intent, often interpreting statements as truthful when they are logically sound or factually aligned.

In opinion- and belief-based datasets such as DEC_OP, CROSS_CULT_DE, and DEFABEL, statements expressing subjective viewpoints or personal convictions are frequently judged as truthful, with reason titles commonly assigned to the *Personal Belief and Values* cluster. This suggests that the model treats the expression of belief itself as evidence of sincerity, independent of whether the belief is intended to mislead. In the personal narrative dataset INTENTION, deceptive inputs are often justified as truthful based on the presence of a sincere tone, with many misclassifications falling under *Honest Tone with No Deception*. This highlights the model's difficulty in detecting deception in subjective or experiential content, where factual verification is not possible. Interestingly, in DEFABEL datasets, a notable number of false negatives are explained using the *Truthful Intent* cluster, suggesting that the model attempts to assess communicative intent, as instructed in the prompts, but lacks a reliable basis for inferring it.

Taken together, these findings suggest that the model equates emotional tone, factual plausibility, and personal expression with truthfulness. In doing so, it overgeneralizes from linguistic cues and fails to capture the deeper pragmatics of deceptive communication.

In the case of false positives—truthful instances incorrectly labeled as deceptive—the most common reasoning cluster is *Misrepresentation of Facts or Science*. This supports our hypothesis that LLMs tend to treat factual accuracy as a proxy for honesty, using it as a primary basis for deception judgments. In OPEN_DOMAIN, a large proportion of false positives are attributed to *Contradictions and Logical Inconsistencies*, with explanations citing inconsistencies in narrative, timeline, or detail. Such cues are often associated with deception in psychological literature, particularly in high-stakes scenarios, where the cognitive load of lying can make it difficult to maintain a coherent story. However, in this context, the model appears to misapply these heuristics, likely due to its exposure to generalized patterns of deceptive behavior during training. As a result, even truthful content that deviates from strict factual or narrative coherence may be incorrectly flagged as deceptive.

When comparing the German and English DEFABEL datasets, we find that the distribution of reasoning for false negatives is quite similar in the belief-based deception datasets compared to other datasets. In the case of false positives, DEFABEL_V2_DE and DEFABEL_V2_EN, which were collected using the same setup, also show comparable patterns. This suggests that the truth-bias behavior does not vary considerably across languages, but rather across datasets.

While the above patterns capture the main quantitative trends, a closer manual inspection of model explanations reveals additional qualitative insights. We find that Mistral rarely engages in explicit reasoning about deceptive intent and instead relies predominantly on judgments of factual accuracy. Even when a statement is factually incorrect, the model often classifies it as truthful if it appears plausible or confidently stated by the author. Table 4 presents a few illustrative examples from the DEFABEL_V2_EN dataset, all of which were labeled as truthful by the model. In the first two instances, the factual content aligns with the author's belief, whereas in the latter two it does not. Nevertheless, the model predicts all of them as truthful, providing rationales that emphasize factual accuracy without critical evaluation. These cases suggest that the model tends to echo or "parrot" the authors' claims rather than verify them, highlighting one of the many inconsistencies in its reasoning process.

Overall, the model appears to conflate factual correctness with honesty and often attempts to justify its predictions by referencing linguistic correlates of deceptive communication, even in cases where these features are not contextually warranted.

6. Conclusion & Future Work

Can deception be detected purely from linguistic cues in text? This question has long guided research in computational linguistics, showing promising results across a variety of datasets. Previous work often relies on the assumption that deception creates cognitive strain, which in turn reveals itself through linguistic markers, suggesting that surface-level textual features can reliably signal deceptive intent. In this paper, we critically reassess these assumptions through the belief-based deception framework.

By introducing the DEFABEL corpus—a multilingual dataset grounded in belief-based deception annotations—we present an alternative framework that isolates deceptive intent from factual correctness. This shift in operationalization allows us to examine deception as a misalignment between belief and expression, irrespective of factual accuracy. Across all variants of DEFABEL, we find that linguistic cues traditionally associated with deception show no significant correlation with deception labels. In similar datasets, these cues exhibit weak and inconsistent correlations. Models that perform well on existing deception datasets, including feature-based classifiers, Transformer models, and

Table 4
Representative examples from the DEFABEL_V2_EN dataset showing author arguments, ground-truth labels, and the model's explanatory rationales.

Argument	True Label	Reason Title	Rationale
Camels store fat in their hump(s) as a source of energy. Where they live means that they may go without food for long periods and long distances. It is not possible for a camel to store water and it is a common misconception that humps contain water.	truthful	No deception detected	The author is providing accurate information about camels, stating that they store fat in their humps as a source of energy and clarifying that it is a common misconception that humps contain water. There is no indication that the author is intentionally stating something they know to be untrue or deliberately misrepresenting their thoughts, feelings, intentions, or actions.
Camels have to store water in their humps because they travel long distances in the desert in very high temperatures and would die of thirst if they were unable to do this. Many people think that the humps contain fat but this is a common misconception!	deceptive	No deception detected	The author is stating a fact about camels and their humps, and there is no evidence to suggest that they are intentionally misrepresenting this information. While it is true that some people may hold the misconception that camel humps contain fat, the author is not one of them and is instead correcting this misconception.
Camels store water in their humps to survive in the desert. While people often think their humps only store fat, the truth is the humps also play a key role in holding water. Camels are known for being able to survive without water for a long time, and this is possible because their humps store water to keep them hydrated. Without the ability to store water in their humps, camels wouldn't be able to go for days without drinking, especially in harsh desert environments. The humps are a special adaptation that helps camels survive in places with little water. This shows that the humps are important for storing water, not just fat, and are crucial for camels living in the desert.	truthful	The text accurately describes the function of camel humps	The author provides a clear and accurate explanation of the role of camel humps in storing water, debunking the common misconception that they only store fat. The text also supports this claim with evidence, such as the camels ability to survive without water for extended periods and the importance of humps in desert survival.
Camels don't store water in their humps. Their humps are full of fat, not water. They use this fat for energy when they can't find food. Camels are good at saving water, but they keep it in their blood, not their humps. They can drink a lot of water at once and stay hydrated for a long time. If humps stored water, they would get smaller when the camel got thirsty. But humps only shrink when the camel uses up its fat.	deceptive	The author is stating accurate information	The author is providing correct information about camels and their humps. They explain that camel humps are not storage for water, but rather contain fat for energy, and that camels store water in their blood. The author also clarifies common misconceptions about camels storing water in their humps and how their humps shrink when they use up their fat, not when they get thirsty.

instruction-tuned LLMs, consistently fail under our belief-based paradigm. Instruction-tuned models in particular display a pronounced truth bias, consistently classifying most statements as truthful regardless of the author's intent. A similar observation has been made by Chuang et al. (2024) showing that LLMs have a bias towards producing factually accurate information in line with scientific reality, despite being prompted

to support an opposing stand. Although CoT prompting can partially reverse this tendency by increasing deceptive classifications, neither prompt variation nor model scaling effectively mitigates the bias. These findings suggest that prior successes in deception detection may stem more from dataset-specific artifacts than from any stable, generalizable signals of deception, which also invite broader reflection on the nature of deception in language and discourse.

Open Question. Why do deception cues vanish in the DEFABEL setting? We hypothesize that the argumentative nature of the task, which explicitly instructs the participants to convince another person, plays a critical role. Deceptive arguments, especially in low-stakes experimental settings, are not spontaneous lies but carefully constructed rhetorical acts. They do not necessarily impose heightened cognitive load, nor do they evoke emotional discomfort; instead, they may rely on the same linguistic strategies used in persuasive writing. As a result, deception in this context may not disrupt language, instead may refine it. This blurs the line between deception and persuasion: the liar and the debater may be linguistically indistinguishable, not because deception leaves no trace, but because argumentation offers tools to mask it.

This interpretation also helps explain why our findings may not directly extend to higher-stakes or emotionally charged forms of deception. In such contexts, the psychological and social pressures surrounding truth and consequence could yield different linguistic dynamics, ones less constrained by rhetorical form and more influenced by affective or cognitive strain.

Our findings thus call for a broader rethinking of deception detection *solely from textual data*, in computational linguistics. Rather than treating deception as a fixed linguistic pattern, we propose viewing it as a context-sensitive, pragmatically constructed phenomenon, rooted in speaker intent and communicative setting. In doing so, we shift the analytical focus from linguistic features to epistemic alignment and communicative goals. A closely related perspective comes from Li and Santos (2012), one of the few studies to explicitly model deception in structured, argumentative settings. Their work simulates social interactions between agents, capturing both the internal reasoning behind deceptive arguments and the communicative strategies used to present them. Deception, in their model, emerges through the intentional construction of arguments designed to align with the expectations of others. Crucially, they find that deception leaves behind detectable patterns in belief manipulation and argument structure. Our findings present a contrast: In real-world rhetorical contexts, such as those represented in our corpus, these patterns may be masked rather than revealed. When deception is embedded in persuasive, fluent argumentation, it may not disrupt linguistic or structural norms at all. This divergence points to a key question for future work: under what conditions do belief-based signals of deception survive, and when are they erased by the demands of genre and discourse?

What if Deception Cannot be Detected? This question, posed in the title of our paper, serves as a final reflection. If deception, specifically in low-stakes contexts, cannot be reliably detected through linguistic cues alone, then the widespread deployment of automated deception detection systems must be approached with extreme caution. Our findings challenge the foundational assumption that textual features can capture the intent to deceive, especially when separated from factual accuracy. This casts doubt on the reliability and ethical defensibility of systems that claim to detect deception in online-communication.

More broadly, our results urge the community to reconsider the framing of deception detection as a solvable classification task. If deception is inherently contextual, strategic, and shaped by interpersonal and cultural dynamics, then purely text-based models, no matter how advanced, may always fall short. Future work must grapple with these limitations, exploring richer representations of belief, context, and communicative intent, or perhaps acknowledging the limits of what such systems can responsibly infer.

Future Work. Our findings point toward a deeper reconceptualization of deception in structured discourse. In argumentative settings, which reward coherence, fluency, and persuasiveness, deception may not manifest through disruption or leakage, but through conformity. Rather than breaking with linguistic norms, deceptive statements may align closely with the rhetorical expectations of the genre, effectively camouflaging intent behind structured language. This echoes Goffman's notion of performance in social interaction, where the speaker's presentation of self is always strategically constructed to fit the frame of the situation (Goffman 1959). Similarly, in the context of mediated communication, Turkle (Turkle 1995) noted that text-based environments allow for the construction of idealized selves, where messages can be carefully "*edited*, *deleted*, and *retouched*". Walther's hyperpersonal model (Walther 1996) likewise emphasizes how low-cue, asynchronous settings enable users to curate their personas in ways that obscure intent. In these contexts, deception becomes less a linguistic anomaly and more a stylized act, indistinguishable from persuasive stance-taking. Taken together, these insights suggest that the absence of detectable cues in our study may not reflect a shortcoming of computational models, but rather the success of deceivers in adapting to the communicative affordances and rhetorical conventions of mediated, structured discourse.

This shift opens up several directions for future research. One is to investigate how deception behaves across different communicative genres—for instance, comparing structured argumentation with more spontaneous, less rhetorically constrained settings. Another is to examine whether belief-expression misalignment becomes more salient under higher-stakes or emotionally charged conditions. Finally, while our results suggest that linguistic cues are weak signals in rhetorical contexts, future work may explore whether deeper models of reasoning and audience design can recover latent traces of deceptive intent. Understanding when, where, and how deception becomes visible remains an open challenge that lies at the intersection of language, intent, and context.

Limitations

This study investigates linguistic cues of deception in low-stakes scenarios, across German and English. While we analyze how previously reported cues behave on selected datasets and evaluate the performance of existing state-of-the-art models, our primary focus is not on optimizing model accuracy. Our aim is to understand how deception manifests linguistically rather than to build the most effective classifier. We acknowledge that tailored models could yield improved performance for specific use cases. However, our study intentionally prioritizes the linguistic exploration of deception over model refinement, leaving further optimization for future work.

The null findings in our linguistic correlation experiments on the DeFaBel datasets should be interpreted with appropriate caution. The absence of statistically significant correlations between linguistic features and deceptive intent may be due to the specific operationalization of the data collection process or the feature sets examined in this

study, rather than the total absence of linguistic cues of deception. Further, deceptive language in argumentative contexts may differ from other forms of deception. However, we do not explore or contrast linguistic patterns of deception in argumentative contexts with those in other low-stakes scenarios, as this is beyond the scope of the current study. This is an area that warrants further exploration, particularly in understanding how deception patterns vary across different discourse genres and communicative goals.

Beyond methodological scope, deception detection research carries societal limitations. Linguistic cues are not universal; they vary with language, culture, and communication style. Models trained on specific datasets may therefore generalize poorly and risk biased or unfair outcomes, particularly for non-native speakers. These limitations highlight the need for fairness evaluation and transparency in future cross-linguistic work.

Ethical Considerations

This study involves the elicitation and analysis of deceptive arguments in a controlled, low-stakes setting. Participants were asked to argue against their own beliefs, with all prompts carefully curated to avoid sensitive or emotionally charged content. These prompts reflect common misconceptions rather than deeply held values, minimizing potential discomfort.

All participants were informed of the nature and goals of the study in advance and gave explicit consent prior to taking part. They were also reminded of their right to withdraw at any time, without consequence. While we collected demographic information to support analysis, we did not collect or publish any personally identifying information. All data used in our analyses remains fully anonymized.

We are aware of the broader ethical implications of deception detection research. While our research aims to deepen the scientific understanding of deceptive intent through linguistic and computational analysis, such models could be misused, for example, to unfairly profile individuals or generate more convincing disinformation. We also recognize that linguistic cues to deception can vary across languages, cultures, and proficiency levels. Automatic systems trained on one dataset may misinterpret the communication styles of non-native speakers or minority groups, leading to biased or unfair predictions. Mislabeling an author's intent could have reputational or social consequences, particularly in sensitive domains such as education, hiring, or law enforcement. We therefore caution against any deployment of such models in consequential settings, particularly where individual judgment or profiling is involved.

At present, automatic deception detection models should not be relied upon for high-stakes decisions. The performance of such systems remains insufficient for practical deployment, and their use in real-world scenarios risks producing incorrect or unjust outcomes. We advocate for the responsible and transparent use of such models, and stress that any deployment should be accompanied by appropriate ethical oversight.

Appendix A: Belief-based Deception Survey

A.1 Corpus I (DeFaBel_V1_De)

The survey for Corpus I was conducted using Google Forms and consisted of two distinct parts: (1) Belief Assessment and (2) Argument Generation.

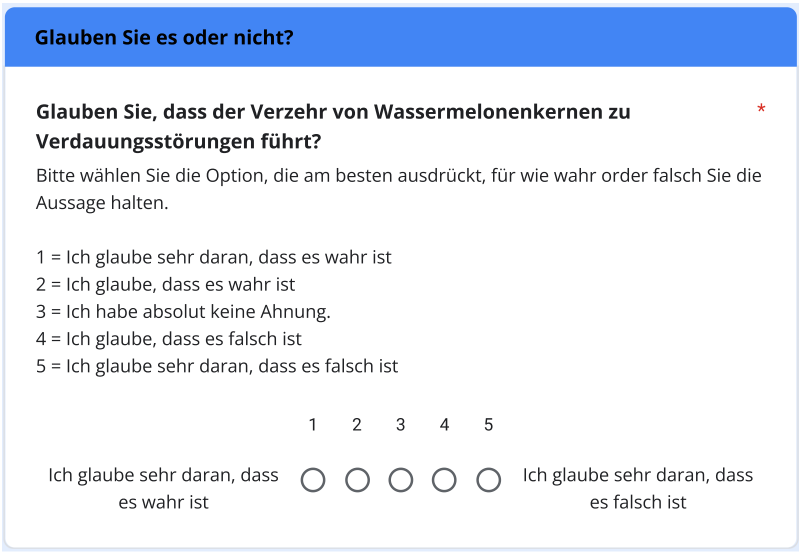


Figure A.1
Belief assessment question as displayed in Google Forms.

Belief Assessment. In this step, participants were asked to indicate their level of agreement with a given statement using a 5-point Likert scale, as illustrated in Figure A.1. To ensure attentiveness, five attention check questions were interspersed throughout the survey.

Argument Generation. Here, participants were prompted to compose convincing arguments in support of a given statement. An example response was provided to set expectations for the task. After completing their written arguments, participants were asked to self-report their familiarity with the topic and their confidence in the persuasiveness of their argument, as shown in Figure A.2. Originally, topic familiarity was recorded on a scale from 0 to 3, and confidence in argument convincingness on a scale from 0 to 4. For consistency in comparative analysis presented in this paper, these scales are normalized to 1–4 and 1–5, respectively.

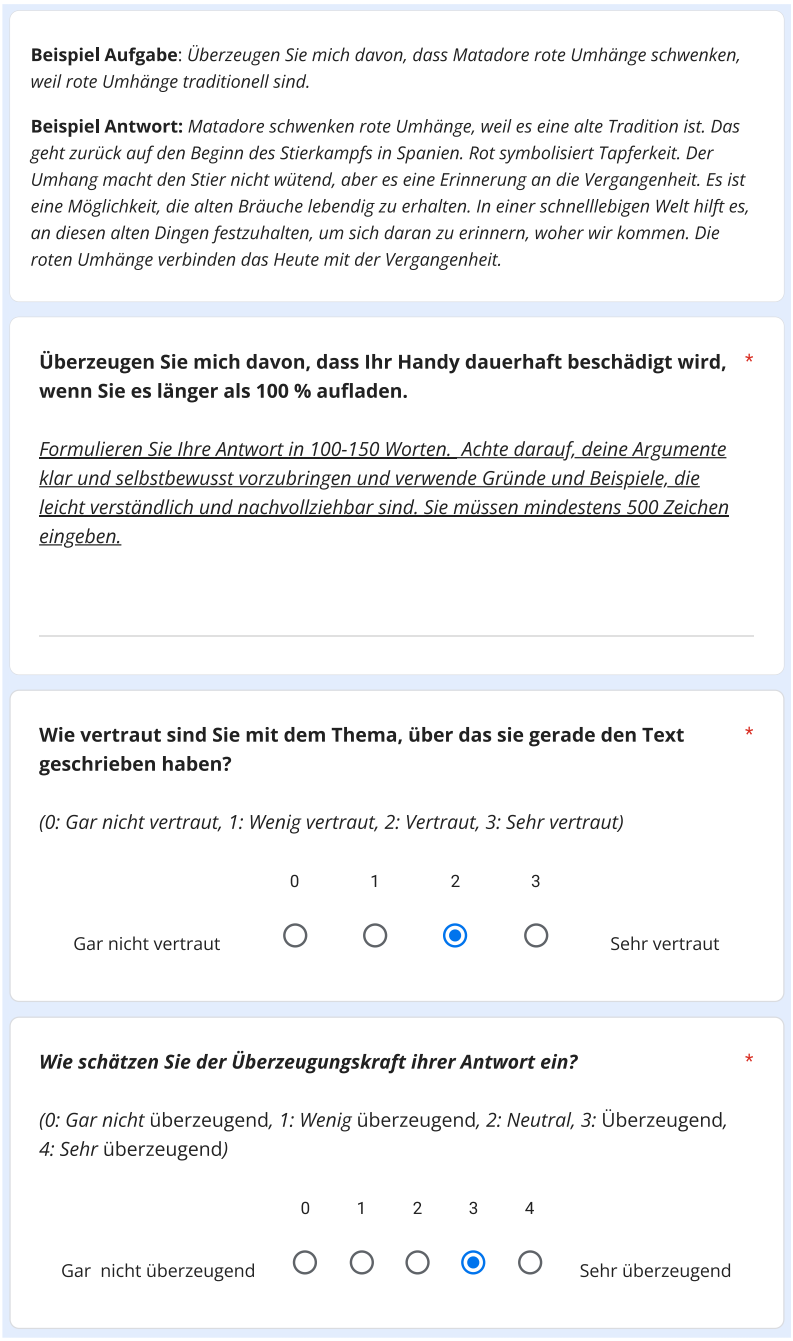


Figure A.2
Argument generation task as displayed in Google Forms.

A.2 Corpus II (DeFaBel_V2_De) & Corpus III (DeFaBel_V2_En)

For the data collection in DeFaBel_V2_De and DeFaBel_V2_En, the survey was implemented using the `Streamlit` package in Python, which is an interactive Web-based application with a more dynamic design. Following are the screenshots (A.3 to A.12) of the annotation interface as implemented in the Streamlit application, illustrating the layout and flow of the tasks.

A.3 Stimuli Design

The stimuli used to generate DeFaBel_V1_De and DeFaBel_V2_De in German differ slightly in their construction. For DeFaBel_V1_De, each item consists of either a factual or a non-factual variant of a given statement. In contrast, DeFaBel_V2_De includes both the factual and the non-factual variants for each statement. For the English version, DeFaBel_V2_En, we translated both variants from DeFaBel_V2_De into English. Tables A.1, A.2, and A.3, present the full list of stimuli used for creating each of the respective corpora.

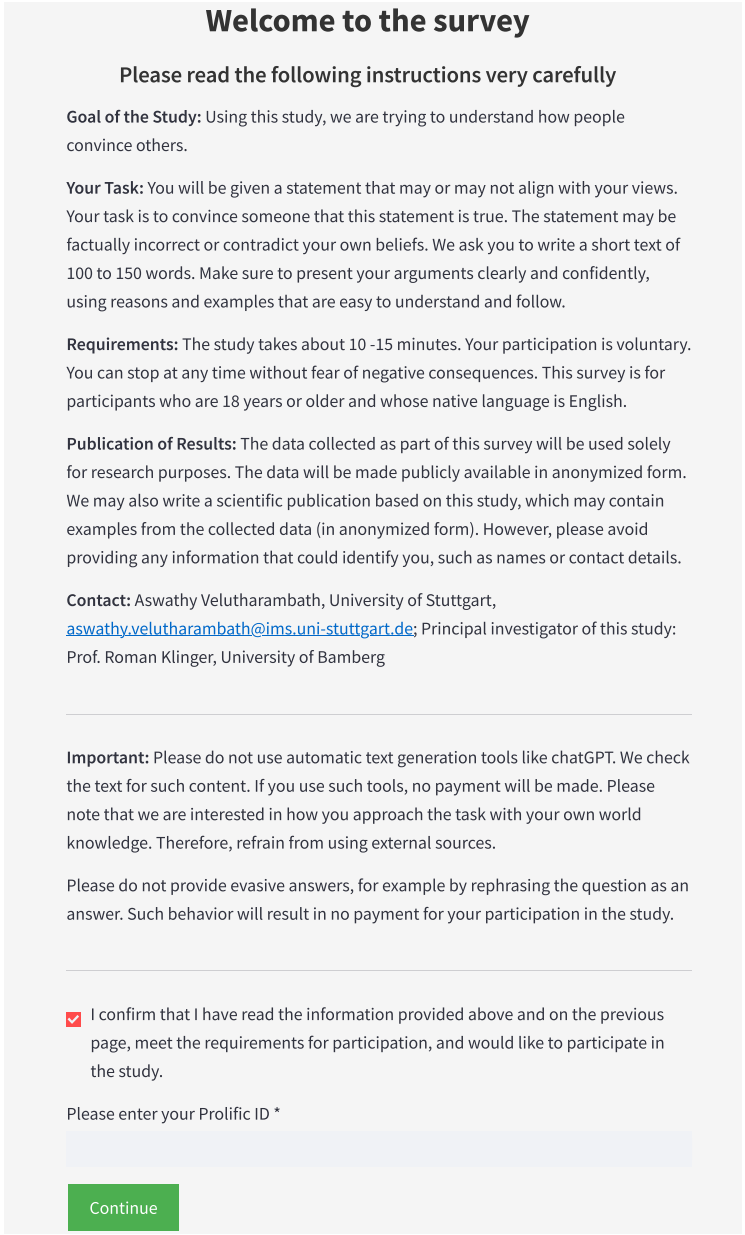


Figure A.3
Task introduction page.

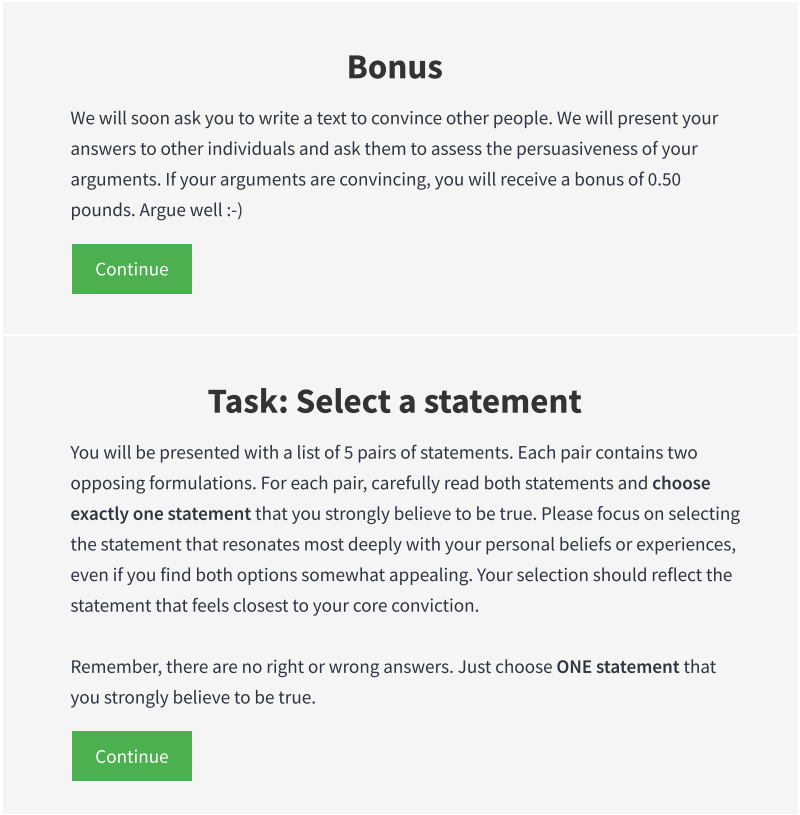


Figure A.4
Top: Bonus announcement page. **Bottom**: Instruction for the statement selection task.



Figure A.5
Statement selection.

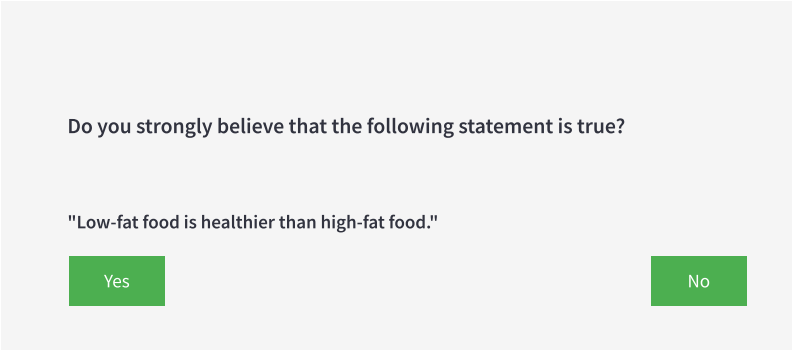


Figure A.6
Belief confirmation.

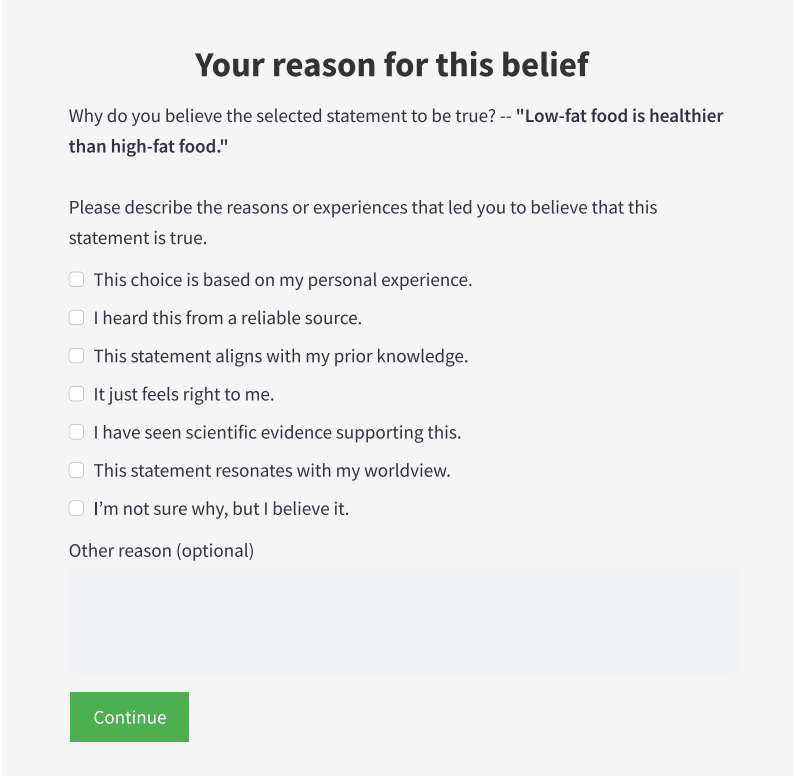


Figure A.7
Belief commitment.

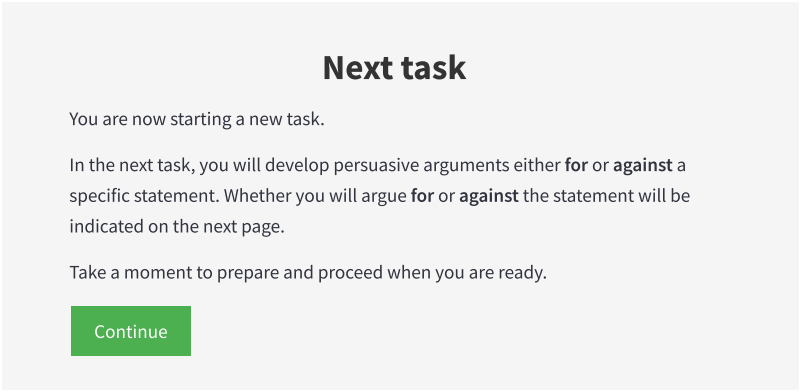


Figure A.8
Task 1 - Introduction.

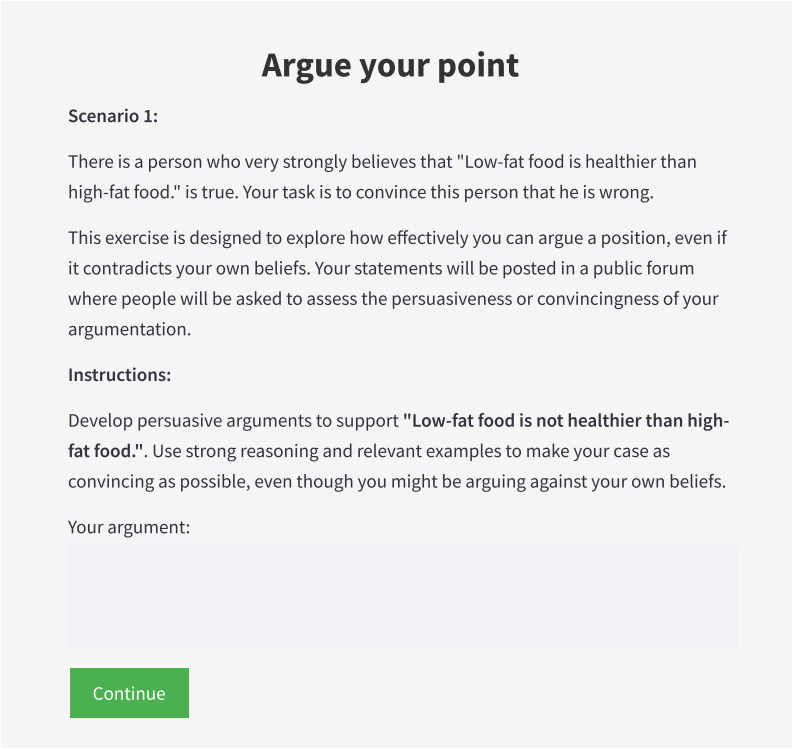


Figure A.9
Task 1.

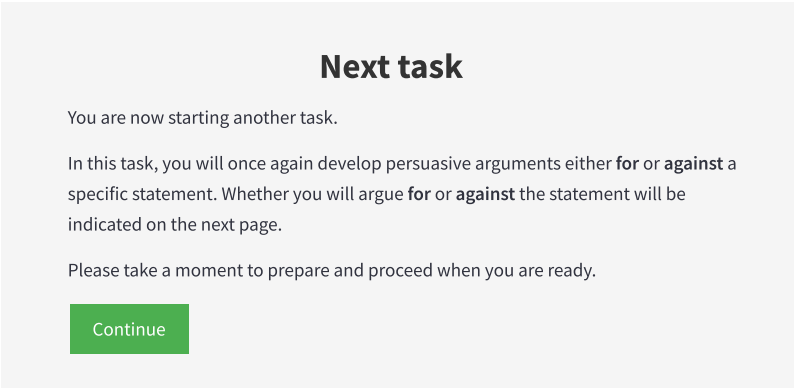


Figure A.10
Task 2 - Introduction.

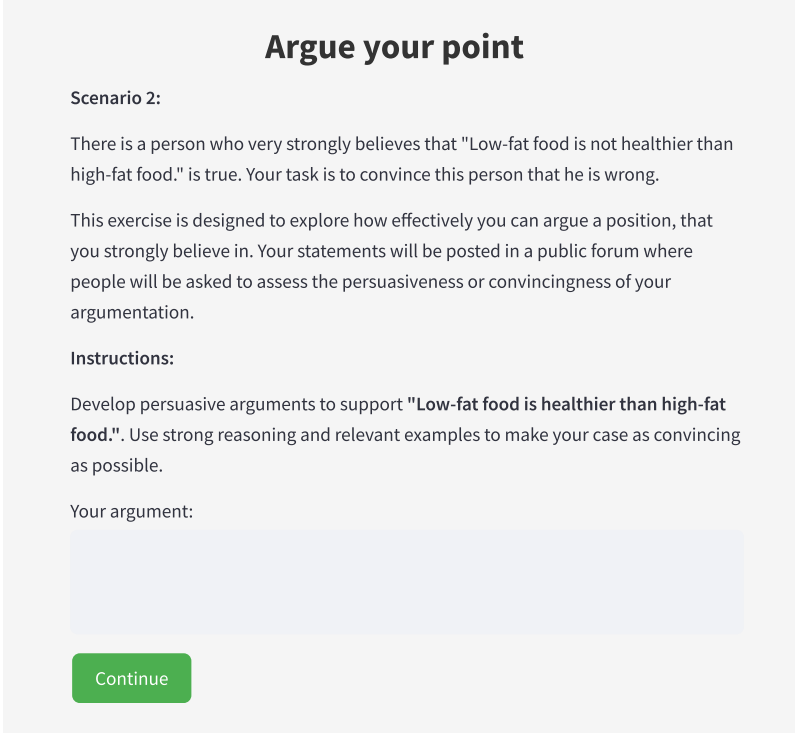


Figure A.11
Task 2.

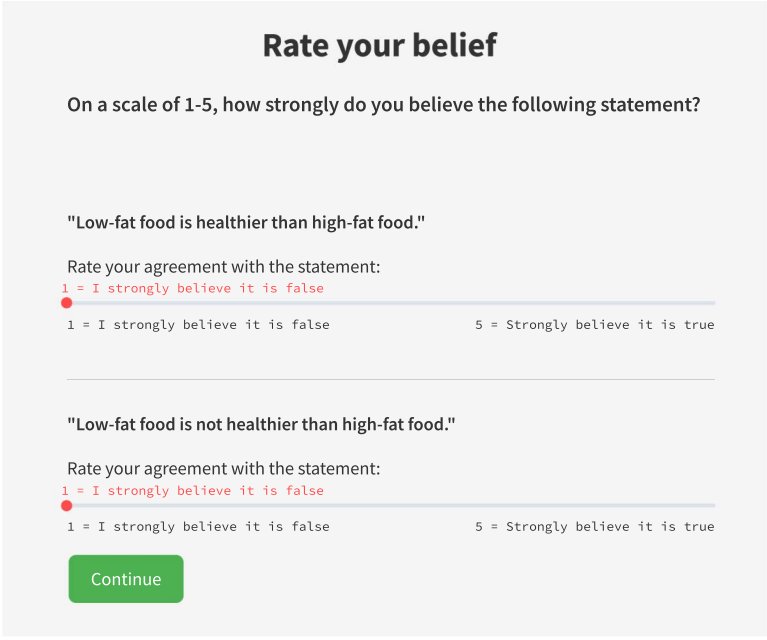


Figure A.12
Belief rating.

Table A.1
Stimuli used for each of the corpora (1–11 of 30).

DEFABEL_V1_DE	DEFABEL_V2_DE	DEFABEL_V2_EN
Der Verzehr von Wassermelonenkernen führt zu Verdauungsstörungen.	Wenn man Wassermelonenkernen isst, passieren sie das Verdauungssystem und werden ausgeschieden. Der Verzehr von Wassermelonenkernen führt zu Verdauungsstörungen.	When watermelon seeds are eaten, they pass through the digestive system and are excreted. The consumption of watermelon seeds leads to indigestion.
Glückskekse haben ihren Ursprung in den Vereinigten Staaten.	Glückskekse haben ihren Ursprung in China. Glückskekse haben ihren Ursprung in den Vereinigten Staaten.	Fortune cookies have their origins in China. Fortune cookies have their origins in the United States.
Der Mensch nutzt die meiste Zeit nur 10% seines Gehirns.	Der Mensch nutzt die meiste Zeit nur 10% seines Gehirns. Das Gehirn des Menschen ist fast immer vollständig aktiv.	Humans use only 10% of their brain most of the time. The human brain is almost always fully active.
Die Sonne erscheint aus dem Weltraum gesehen weiß.	Die Sonne erscheint aus dem Weltraum gesehen weiß. Die Sonne erscheint aus dem Weltraum gesehen gelb.	The sun appears white when viewed from space. The sun appears yellow when viewed from space.
Wenn man einen Regenwurm durchschneidet, entstehen zwei Regenwürmer.	Wenn man einen Regenwurm durchschneidet, stirbt er sehr wahrscheinlich. Wenn man einen Regenwurm durchschneidet, entstehen zwei Regenwürmer.	If you cut through an earthworm, you get two earthworms. If you cut an earthworm, it is very likely to die.
In der Bibel wird erwähnt, welche verbotene Frucht Adam und Eva im Garten Eden aßen.	In der Bibel wird nicht erwähnt, welche verbotene Frucht Adam und Eva im Garten Eden aßen. In der Bibel wird erwähnt, welche verbotene Frucht Adam und Eva im Garten Eden aßen.	The Bible mentions the forbidden fruit that Adam and Eve ate in the Garden of Eden. The Bible does not mention which forbidden fruit Adam and Eve ate in the Garden of Eden.
In kleinen Dosen ist Glutamat (ein Geschmacksverstärker) schädlich für den Menschen.	In kleinen Dosen ist Glutamat (ein Geschmacksverstärker) schädlich für den Menschen. Es gibt keine eindeutigen wissenschaftlichen Beweise dafür, dass Glutamat (ein Geschmacksverstärker) in kleinen Dosen für den Menschen schädlich ist.	There is no clear scientific evidence that glutamate (a flavor enhancer) is harmful to humans in small doses. In small doses, glutamate (a flavor enhancer) is harmful to humans.
Man sollte nicht zu nah vor dem Fernseher sitzen, weil das zu einer vorübergehenden Ermüdung der Augen führen kann.	Man sollte nicht zu nah vor dem Fernseher sitzen, weil das zu einer vorübergehenden Ermüdung der Augen führen kann. Man sollte nicht zu nah am Fernseher sitzen, da dies zur Erblindung führen kann.	You should not sit too close to the TV as this can cause temporary eye fatigue. You should not sit too close to the television as this can lead to blindness.
Glutenfreies Brot hat gesundheitliche Vorteile.	Glutenfreies Brot hat gesundheitliche Vorteile. Glutenfreies Brot ist nur besser, wenn man glutenintolerant ist.	Gluten-free bread is only better if you are gluten intolerant. Gluten-free bread has health benefits.
Mäßiger Weinkonsum ist nachweislich gut für die Gesundheit.	Es gibt keine klaren wissenschaftlichen Beweise dafür, dass mäßiger Weinkonsum gut für die Gesundheit ist. Mäßiger Weinkonsum ist nachweislich gut für die Gesundheit.	There is no clear scientific evidence that moderate wine consumption is good for your health. Moderate wine consumption is proven to be good for your health.
Luft ist dichter als Wasserdampf.	Luft ist dichter als Wasserdampf. Wasserdampf ist dichter als Luft.	Air is denser than water vapor. Water vapor is denser than air.

Table A.2
Stimuli used for each of the corpora (12–24 of 30).

DEFABEL_V1_DE	DEFABEL_V2_DE	DEFABEL_V2_EN
Bio-Lebensmittel sind gesünder als konventionelle Lebensmittel.	Bio-Lebensmittel sind gesünder als konventionelle Lebensmittel. Konventionelle Lebensmittel sind genauso gesund wie Bio-Lebensmittel.	Organic food is healthier than conventional food. Conventional food is just as healthy as organic food.
In der ökologischen Landwirtschaft werden Pestizide verwendet.	In der ökologischen Landwirtschaft werden Pestizide verwendet. In der ökologischen Landwirtschaft sind Pestizide ausschließlich verboten.	Pesticides are strictly prohibited in organic farming. Pesticides are used in organic farming.
Das Kauen von Kaugummi ist in Singapur illegal.	Das Kauen von Kaugummi ist in Singapur illegal. In Singapur ist es nur illegal, Kaugummi zu importieren und zu verkaufen, nicht es zu kauen.	Chewing gum is illegal in Singapore. In Singapore, it is only illegal to import and sell chewing gum, not to chew it.
Mehr Amerikaner besuchen eine Hochschule oder Universität als Europäer.	Es besuchen mehr Amerikaner als Europäer eine Hochschule oder Universität. Es besuchen mehr Europäer als Amerikaner eine Hochschule oder Universität.	More Europeans attend college or university than Americans. More Americans than Europeans attend college or university.
Atheisten schreiben häufiger 'Xmas' als Christen, die eher 'Weihnachten' schreiben.	Die Verwendung von 'Xmas' oder 'Weihnachten' hat nichts mit Glauben zu tun, sondern ist lediglich eine Abkürzung. Atheisten schreiben häufiger 'Xmas' als Christen, die eher 'Weihnachten' schreiben.	The use of 'Xmas' or 'Christmas' has nothing to do with faith, but is merely an abbreviation. Atheists are more likely to write 'Xmas' than Christians, who are more likely to write 'Christmas'.
Der Verzehr von Brokkoli kann Ihre Lebenserwartung um viele Jahrzehnte verlängern.	Es wurde wissenschaftlich nicht nachgewiesen, dass der Verzehr von Gemüse Ihre Lebenserwartung um viele Jahrzehnte verlängert. Der Verzehr von Brokkoli kann Ihre Lebenserwartung um viele Jahrzehnte verlängern.	It has not been scientifically proven that eating vegetables extends your life expectancy by many decades. Eating broccoli can extend your life expectancy by many decades.
Es gibt auf der Erde einen Ort, an dem immer Nacht ist.	Es gibt auf der Erde einen Ort, an dem immer Nacht ist. Die Sonne scheint nie am Südpol.	There is a place on earth where it is always night. The sun never shines at the South Pole.
Manche Menschen können im Schlaf lernen.	Manche Menschen können im Schlaf lernen. Es gibt keine wissenschaftlichen Beweise dafür, dass Menschen im Schlaf lernen können.	Some people can learn in their sleep. There is no scientific evidence that people can learn in their sleep.
Es gibt Impfstoffe, die dauerhaft deine DNA verändern können.	Impfstoffe haben keine dauerhaften Auswirkungen auf die DNA. Es gibt Impfstoffe, die dauerhaft die DNA verändern können.	Vaccines have no lasting effects on DNA. There are vaccines that can permanently alter the DNA.
Hitler hat mit dem Bau der Autobahnen begonnen.	Der Bau der Autobahnen begann bereits vor Hitlers Diktatur. Hitler hat mit dem Bau der Autobahnen begonnen.	Hitler started building the autobahns. The construction of the autobahns began even before Hitler's dictatorship.
Haie können Krebs bekommen.	Haie können Krebs bekommen. Haie bekommen niemals Krebs, weil ihr Knorpel als Krebsschutz wirkt.	Sharks can get cancer. Sharks never get cancer because their cartilage acts as cancer protection.
Kamele speichern Fett in ihren Höckern.	Kamele speichern Wasser in ihren Höckern. Kamele speichern Fett in ihren Höckern.	Camels store water in their humps. Camels store fat in their humps.
Die Erde ist im Januar der Sonne am nächsten.	Die Erde ist im Juli der Sonne am nächsten. Die Erde ist im Januar der Sonne am nächsten.	The earth is closest to the sun in July. The earth is closest to the sun in January.

Table A.3
Stimuli used for each of the corpora (25–30 of 30).

DEFABEL_V1_DE	DEFABEL_V2_DE	DEFABEL_V2_EN
Die Nutzung von Handys an Tankstellen kann eine Explosion verursachen.	Die Nutzung von Handys an Tankstellen kann eine Explosion verursachen. Die Nutzung von Handys an Tankstellen ist untersagt, um mögliche Sicherheitsrisiken durch Ablenkung zu minimieren.	The use of cell phones at filling stations is prohibited in order to minimize possible safety risks due to distraction. The use of cell phones at gas stations can cause an explosion.
Ihr Handy wird dauerhaft beschädigt, wenn Sie länger als 100 % aufladen.	Ihr Handy wird dauerhaft beschädigt, wenn Sie es länger als 100 % aufladen. Ihr Handy kann problemlos länger als bis zu 100 % aufgeladen werden.	Your cell phone can easily be charged for longer than 100%. Your cell phone will be permanently damaged if you charge it for longer than 100%.
Affen können genauso gut addieren wie Menschen.	Kein Tier kann genauso gut wie der Mensch Zahlen addieren. Affen können genauso gut addieren wie Menschen.	No animal can add numbers as well as humans. Monkeys can add just as well as humans.
Es ist in Deutschland illegal, barfuß Auto zu fahren.	Es ist in Deutschland illegal, barfuß Auto zu fahren. Es gibt keine gesetzliche Einschränkung in Deutschland, barfuß Auto zu fahren.	It is illegal to drive barefoot in Germany. There is no legal restriction on driving barefoot in Germany.
Fettarmes Essen ist gesünder als fettreiches Essen.	Fettarmes Essen ist gesünder als fettreiches Essen. Fettarmes Essen ist nicht gesünder als fettreiche Essen.	Low-fat food is healthier than high-fat food. Low-fat food is not healthier than high-fat food.
Tigerhaie untergehen, wenn sie aufhören zu schwimmen.	Tigerhaie gehen unter, wenn sie aufhören zu schwimmen. Tigerhaie treiben auf dem Wasser, wenn sie aufhören zu schwimmen.	Tiger sharks float on the water when they stop swimming. Tiger sharks sink when they stop swimming.

Appendix B: Experiment Details

B.1 Linguistic Cues as Features for Deception Detection

To analyze deception in language, we extract a rich set of linguistic cues grounded in prior psycholinguistic and computational studies. These features span multiple dimensions, including affective tone, complexity, specificity, and syntactic structure, and are designed to capture both subtle and explicit markers of deceptive behavior. Table B.1 provides an overview of the feature categories, the specific cues used, and how each is operationalized.

B.2 Linguistic Correlation Analysis

Tables B.2 and Table B.3 show the point-biserial correlation values and the lower bound and upper bound of the 95% confidence interval; categories with correlations considered significant (p-value < 0.05) are marked with a * symbol.

Table B.1
Linguistic cues and their operationalization.

Linguistic cues	Features	Operationalization
Authenticity	Authentic	Authenticity score (LIWC)
Affect	Tone, Affect, tone_pos, tone_neg	Sentiment score
	emotion, emo_pos, emo_neg	Emotion scores
	emo_anx, emo_anger, emo_sad	Anxiety, anger, sadness scores
	positive, negative, neutral	Positive, negative, neutral scores
Quantity	WPS, word_count, sentence_count	Words per sentence, word and sentence counts
	syllable_count, avg_syllables_per_word	Syllable count, average syllables per word
	verb, VERB	Verb frequency
Distancing	ppron, ipron, negate	Pronouns and negations
	i, we, you, shehe, they	Personal pronouns
	self_reference, other_reference	Self and other references
Activation	adverb	Frequency of adverbs
Complexity	adj	Frequency of adjectives
	differ	Exclusive words
	flesch_kincaid_grade	Flesch–Kincaid readability
	gunning_fog	Gunning–Fog readability
Specificity	quantity	Frequency of quantifiers
	Cognition, cogproc	Cognitive processing words
	insight, cause, memory	Insight, causation, memory words
	Perception, attention	Perception and attention words
	motion, space, time	Motion, space, and time references
	visual, auditory, feeling	Visual, auditory, feeling references
	PROPN	Proper noun frequency
	people_entities, temporal_entities	Entity recognition: people, temporal
	spatial_entities, quantity_entities	Entity recognition: spatial, quantity
	perceptual_details	Perceptual detail markers
	contextual_embedding	Contextual embedding score
	reality_monitoring	Reality monitoring features
Uncertainty	discrep, tentat, certitude	Discrepancy, tentative, certitude markers
Diversity	lexical_diversity	Lexical diversity (unique tokens / tokens)
	difficult_words	Difficult word count

Table B.2
Point-biserial correlations by dataset (German).

	DeFaBel_V1_De	DeFaBel_V2_De
Categories	correlation[CI]	correlation[CI]
Authentic	−0.01 [−0.08, 0.05]	−0.05 [−0.11, 0.01]
Tone	−0.00 [−0.06, 0.06]	−0.03 [−0.09, 0.03]
affect	0.05 [−0.01, 0.12]	0.05 [−0.01, 0.11]
positive	0.03 [−0.03, 0.09]	−0.02 [−0.08, 0.05]
negative	0.05 [−0.01, 0.11]	0.04 [−0.03, 0.10]
neutral	−0.06 [−0.12, 0.01]	−0.02 [−0.08, 0.04]
posemo	0.02 [−0.04, 0.08]	0.02 [−0.05, 0.08]
negemo	0.06 [−0.00, 0.12]	0.05 [−0.02, 0.11]
anx	0.07 [0.01, 0.13]	0.08 [0.02, 0.14]
anger	0.07 [0.01, 0.13]	0.04 [−0.03, 0.10]
sad	−0.03 [−0.09, 0.03]	−0.00 [−0.07, 0.06]
WPS	−0.02 [−0.08, 0.04]	−0.00 [−0.07, 0.06]
verb	0.01 [−0.05, 0.07]	0.01 [−0.05, 0.08]
word_count	−0.05 [−0.11, 0.01]	−0.08 [−0.14, −0.02]
sentence_count	−0.02 [−0.08, 0.04]	−0.09 [−0.15, −0.02]
syllable_count	−0.05 [−0.11, 0.01]	−0.08 [−0.14, −0.02]
avg_syllables_per_word	0.01 [−0.05, 0.07]	−0.01 [−0.07, 0.05]
VERB	0.01 [−0.05, 0.07]	0.00 [−0.06, 0.06]
ppron	−0.05 [−0.11, 0.01]	0.03 [−0.03, 0.09]
i	−0.01 [−0.07, 0.05]	−0.00 [−0.07, 0.06]
we	−0.02 [−0.08, 0.05]	−0.00 [−0.07, 0.06]
you_sing	−0.02 [−0.08, 0.05]	−0.01 [−0.07, 0.05]
you_plur	−0.02 [−0.08, 0.04]	0.01 [−0.05, 0.08]
shehe	−0.03 [−0.09, 0.03]	0.05 [−0.02, 0.11]
they	−0.00 [−0.07, 0.06]	0.03 [−0.03, 0.09]
ipron	0.02 [−0.04, 0.08]	−0.04 [−0.10, 0.03]
negate	−0.02 [−0.09, 0.04]	0.05 [−0.01, 0.11]
self_reference	−0.02 [−0.08, 0.04]	−0.00 [−0.07, 0.06]
other_reference	−0.02 [−0.08, 0.04]	0.04 [−0.02, 0.10]
adverb	−0.06 [−0.12, 0.00]	−0.04 [−0.10, 0.03]
adj	0.06 [−0.01, 0.12]	0.00 [−0.06, 0.06]
differ	0.01 [−0.05, 0.07]	−0.03 [−0.10, 0.03]
flesch_kincaid_grade	−0.02 [−0.08, 0.04]	0.00 [−0.06, 0.07]
gunning_fog	−0.02 [−0.08, 0.04]	0.01 [−0.05, 0.07]
cogproc	0.01 [−0.05, 0.07]	−0.00 [−0.06, 0.06]
insight	−0.00 [−0.06, 0.06]	0.01 [−0.05, 0.07]
cause	−0.03 [−0.09, 0.03]	0.06 [−0.00, 0.12]
motion	−0.05 [−0.11, 0.02]	0.02 [−0.04, 0.08]
space	−0.01 [−0.07, 0.05]	−0.02 [−0.08, 0.04]
see	0.09 [0.03, 0.15]	−0.02 [−0.09, 0.04]
hear	−0.04 [−0.10, 0.02]	−0.02 [−0.08, 0.04]
feel	−0.00 [−0.06, 0.06]	0.02 [−0.04, 0.09]
percept	0.08 [0.02, 0.14]	−0.00 [−0.07, 0.06]
time	−0.02 [−0.08, 0.04]	−0.05 [−0.12, 0.01]
PROPN	−0.01 [−0.07, 0.05]	−0.03 [−0.10, 0.03]
people_entities	0.07 [0.00, 0.13]	−0.04 [−0.10, 0.03]
spatial_entities	0.02 [−0.05, 0.08]	−0.05 [−0.11, 0.01]
perceptual_details	0.07 [0.01, 0.13]	−0.01 [−0.08, 0.05]
contextual_embedding	−0.03 [−0.09, 0.03]	−0.04 [−0.10, 0.02]
reality_monitoring	0.03 [−0.03, 0.09]	−0.01 [−0.08, 0.05]
discrep	−0.02 [−0.08, 0.04]	−0.02 [−0.09, 0.04]
tentat	0.05 [−0.01, 0.11]	−0.05 [−0.11, 0.01]
certain	−0.03 [−0.09, 0.03]	0.03 [−0.04, 0.09]
lexical_diversity	0.03 [−0.03, 0.09]	0.05 [−0.02, 0.11]
difficult_words	−0.06 [−0.12, 0.00]	−0.08 [−0.14, −0.02]

Table B.3
Point-biserial correlations by dataset (English).

	DEFABEL_EN	CROSS_CULT_DE	DEC_OP	INTENTION	OPEN_DOMAIN
Categories	correlation[CI]	correlation[CI]	correlation[CI]	correlation[CI]	correlation[CI]
Authentic	−0.05 [−0.12, 0.02]	0.28 [0.22, 0.33]*	0.22 [0.18, 0.26]*	−0.04 [−0.08, 0.01]	−0.04 [−0.07, −0.02]*
Tone	−0.01 [−0.08, 0.06]	−0.01 [−0.07, 0.04]	0.14 [0.10, 0.18]*	0.07 [0.03, 0.12]	0.01 [−0.02, 0.03]
Affect	0.02 [−0.05, 0.09]	−0.09 [−0.15, −0.03]	−0.05 [−0.09, −0.01]	0.05 [0.00, 0.10]	−0.00 [−0.03, 0.02]
tone_pos	−0.00 [−0.07, 0.07]	−0.05 [−0.11, 0.01]	0.13 [0.09, 0.17]*	0.05 [0.00, 0.10]	0.02 [−0.00, 0.04]
tone_neg	0.03 [−0.04, 0.10]	−0.04 [−0.10, 0.02]	−0.14 [−0.18, −0.10]*	0.01 [−0.04, 0.05]	−0.03 [−0.06, −0.01]
emotion	−0.02 [−0.09, 0.05]	−0.08 [−0.14, −0.03]	0.08 [0.04, 0.12]*	0.07 [0.02, 0.12]	0.02 [−0.00, 0.04]
emo_pos	−0.03 [−0.10, 0.04]	−0.09 [−0.15, −0.04]	0.06 [0.02, 0.10]	0.06 [0.01, 0.11]	0.04 [0.02, 0.06]*
emo_neg	−0.00 [−0.07, 0.07]	−0.00 [−0.06, 0.05]	0.03 [−0.01, 0.07]	0.04 [−0.01, 0.09]	−0.02 [−0.04, 0.00]
emo_anx	0.04 [−0.03, 0.11]	0.06 [−0.00, 0.11]	0.02 [−0.02, 0.06]	0.04 [−0.01, 0.09]	−0.01 [−0.04, 0.01]
emo_anger	−0.02 [−0.09, 0.04]	−0.05 [−0.11, 0.01]	0.02 [−0.02, 0.06]	−0.01 [−0.05, 0.04]	−0.03 [−0.05, −0.00]
emo_sad	0.02 [−0.05, 0.09]	0.02 [−0.04, 0.07]	0.04 [0.01, 0.08]	0.01 [−0.04, 0.06]	−0.01 [−0.04, 0.01]
positive	0.00 [−0.07, 0.07]	−0.03 [−0.08, 0.03]	0.16 [0.12, 0.20]*	0.10 [0.05, 0.15]*	0.02 [−0.00, 0.05]
negative	0.04 [−0.03, 0.11]	−0.01 [−0.07, 0.04]	−0.22 [−0.26, −0.18]*	0.01 [−0.04, 0.06]	−0.04 [−0.07, −0.02]*
neutral	−0.03 [−0.10, 0.04]	0.04 [−0.02, 0.10]	0.09 [0.05, 0.13]*	−0.10 [−0.15, −0.05]*	0.01 [−0.02, 0.03]
WPS	−0.00 [−0.07, 0.07]	0.07 [0.01, 0.12]	0.22 [0.18, 0.26]*	0.07 [0.02, 0.12]	−0.02 [−0.04, 0.00]
verb	−0.00 [−0.07, 0.06]	−0.06 [−0.12, −0.00]	−0.07 [−0.10, −0.03]	0.17 [0.13, 0.22]*	−0.01 [−0.03, 0.02]
word_count	−0.09 [−0.16, −0.02]	0.22 [0.16, 0.27]*	0.25 [0.21, 0.29]*	0.11 [0.06, 0.15]*	−0.02 [−0.04, 0.00]
sentence_count	−0.08 [−0.15, −0.01]	0.14 [0.09, 0.20]*	0.08 [0.04, 0.12]*	0.05 [−0.00, 0.09]	−0.00 [−0.02, 0.02]
syllable_count	−0.08 [−0.15, −0.01]	0.20 [0.15, 0.26]*	0.26 [0.23, 0.30]*	0.10 [0.05, 0.15]*	−0.00 [−0.03, 0.02]
avg_syllables_per_word	0.01 [−0.05, 0.08]	−0.03 [−0.08, 0.03]	0.09 [0.05, 0.12]*	−0.02 [−0.06, 0.03]	0.04 [0.02, 0.07]*
VERB	0.07 [−0.00, 0.14]	−0.06 [−0.11, −0.00]	−0.01 [−0.05, 0.03]	0.01 [−0.03, 0.06]	−0.04 [−0.06, −0.02]*
ppron	0.02 [−0.05, 0.09]	−0.00 [−0.06, 0.06]	0.04 [0.00, 0.08]	0.12 [0.08, 0.17]*	−0.06 [−0.08, −0.04]*
i	−0.04 [−0.11, 0.02]	0.13 [0.07, 0.19]*	0.30 [0.27, 0.34]*	0.12 [0.07, 0.17]*	−0.04 [−0.06, −0.01]
we	−0.01 [−0.08, 0.06]	0.13 [0.07, 0.18]*	−0.15 [−0.19, −0.12]*	−0.05 [−0.10, −0.00]	0.00 [−0.02, 0.03]
you	0.03 [−0.04, 0.10]	−0.05 [−0.11, 0.01]	−0.07 [−0.11, −0.03]*	0.01 [−0.04, 0.06]	−0.05 [−0.07, −0.03]*
shehe	−0.02 [−0.09, 0.05]	−0.15 [−0.21, −0.10]*	−0.00 [−0.04, 0.04]	0.06 [0.01, 0.11]	−0.04 [−0.06, −0.01]
they	0.04 [−0.03, 0.11]	−0.08 [−0.14, −0.02]	−0.08 [−0.11, −0.04]*	0.04 [−0.01, 0.09]	−0.02 [−0.04, 0.00]
ipron	−0.05 [−0.11, 0.02]	0.01 [−0.05, 0.06]	0.04 [−0.00, 0.08]	0.14 [0.09, 0.19]*	−0.04 [−0.06, −0.02]
negate	−0.03 [−0.10, 0.04]	−0.13 [−0.18, −0.07]*	−0.10 [−0.14, −0.06]*	0.11 [0.06, 0.16]*	−0.09 [−0.12, −0.07]*
self_reference	−0.03 [−0.10, 0.04]	0.17 [0.12, 0.23]*	0.15 [0.11, 0.19]*	0.09 [0.04, 0.14]*	−0.04 [−0.06, −0.01]
other_reference	0.05 [−0.02, 0.12]	−0.20 [−0.26, −0.15]*	−0.10 [−0.14, −0.07]*	0.07 [0.03, 0.12]	−0.06 [−0.09, −0.04]*
adverb	0.02 [−0.05, 0.09]	−0.03 [−0.09, 0.02]	0.03 [−0.01, 0.07]	0.13 [0.09, 0.18]*	−0.07 [−0.09, −0.05]*
adj	0.03 [−0.04, 0.10]	0.09 [0.03, 0.15]	−0.00 [−0.04, 0.04]	0.06 [0.01, 0.11]	0.03 [0.00, 0.05]
differ	−0.02 [−0.09, 0.04]	0.11 [0.05, 0.17]*	0.10 [0.06, 0.14]*	0.02 [−0.03, 0.07]	−0.01 [−0.03, 0.01]
flesch_kincaid_grade	0.01 [−0.06, 0.07]	0.09 [0.03, 0.14]	0.23 [0.20, 0.27]*	0.07 [0.02, 0.11]	0.02 [−0.00, 0.05]
gunning_fog	0.01 [−0.06, 0.08]	0.09 [0.03, 0.14]	0.21 [0.18, 0.25]*	0.05 [0.00, 0.10]	0.02 [−0.00, 0.04]
quantity	0.01 [−0.06, 0.08]	0.08 [0.02, 0.13]	0.06 [0.02, 0.10]	0.03 [−0.01, 0.08]	−0.00 [−0.02, 0.02]
Cognition	0.02 [−0.05, 0.09]	0.02 [−0.04, 0.07]	0.07 [0.03, 0.11]*	0.14 [0.09, 0.19]*	−0.09 [−0.12, −0.07]*
cogproc	0.04 [−0.03, 0.11]	0.05 [−0.01, 0.10]	0.12 [0.08, 0.16]*	0.12 [0.08, 0.17]*	−0.02 [−0.04, 0.00]
insight	0.00 [−0.07, 0.07]	0.10 [0.05, 0.16]*	0.25 [0.21, 0.29]*	0.10 [0.05, 0.15]*	−0.03 [−0.05, −0.00]
cause	0.07 [−0.00, 0.14]	−0.03 [−0.09, 0.02]	−0.06 [−0.10, −0.02]	0.05 [0.00, 0.10]	−0.01 [−0.04, 0.01]
memory	−0.09 [−0.16, −0.02]	0.06 [−0.00, 0.11]	−0.01 [−0.05, 0.03]	0.03 [−0.02, 0.07]	−0.01 [−0.03, 0.01]
Perception	0.01 [−0.06, 0.08]	0.07 [0.01, 0.12]	−0.01 [−0.05, 0.03]	−0.10 [−0.14, −0.05]*	0.01 [−0.01, 0.03]
attention	0.00 [−0.07, 0.07]	0.03 [−0.03, 0.08]	0.08 [0.04, 0.11]*	−0.02 [−0.07, 0.03]	−0.00 [−0.03, 0.02]
motion	−0.01 [−0.08, 0.06]	0.02 [−0.03, 0.08]	−0.07 [−0.11, −0.04]*	−0.09 [−0.14, −0.04]*	−0.02 [−0.05, 0.00]
space	0.01 [−0.06, 0.08]	0.07 [0.01, 0.13]	−0.05 [−0.09, −0.01]	−0.10 [−0.15, −0.05]*	0.01 [−0.01, 0.04]
visual	0.01 [−0.06, 0.08]	−0.04 [−0.09, 0.02]	0.05 [0.01, 0.09]	0.06 [0.01, 0.11]	0.01 [−0.02, 0.03]
auditory	−0.03 [−0.10, 0.04]	0.00 [−0.06, 0.06]	0.02 [−0.02, 0.06]	0.01 [−0.04, 0.06]	0.01 [−0.02, 0.03]
feeling	−0.01 [−0.08, 0.06]	0.02 [−0.04, 0.07]	0.12 [0.08, 0.16]*	0.02 [−0.02, 0.07]	0.04 [0.01, 0.06]
time	−0.02 [−0.09, 0.05]	0.02 [−0.03, 0.08]	−0.02 [−0.06, 0.02]	−0.02 [−0.07, 0.03]	−0.03 [−0.06, −0.01]
PROPN	−0.00 [−0.07, 0.06]	−0.01 [−0.07, 0.05]	−0.15 [−0.19, −0.11]*	0.05 [−0.00, 0.10]	0.04 [0.02, 0.07]*
people_entities	−0.03 [−0.10, 0.03]	0.01 [−0.05, 0.07]	−0.08 [−0.12, −0.04]*	0.13 [0.08, 0.18]*	0.02 [0.00, 0.05]
temporal_entities	−0.02 [−0.08, 0.05]	0.15 [0.09, 0.20]*	0.02 [−0.02, 0.06]	0.02 [−0.03, 0.07]	0.01 [−0.01, 0.03]
spatial_entities	−0.00 [−0.07, 0.07]	0.09 [0.03, 0.14]	0.04 [−0.00, 0.08]	0.07 [0.02, 0.12]	0.02 [−0.01, 0.04]
quantity_entities	−0.05 [−0.12, 0.02]	0.08 [0.03, 0.14]	0.03 [−0.01, 0.07]	−0.06 [−0.10, −0.01]	−0.01 [−0.03, 0.02]
perceptual_details	0.00 [−0.07, 0.07]	0.00 [−0.05, 0.06]	0.14 [0.11, 0.18]*	0.04 [−0.01, 0.09]	0.02 [−0.00, 0.04]
contextual_embedding	−0.01 [−0.08, 0.06]	0.06 [0.01, 0.12]	−0.07 [−0.11, −0.03]*	−0.11 [−0.15, −0.06]*	−0.02 [−0.04, 0.01]
reality_monitoring	−0.01 [−0.08, 0.06]	−0.02 [−0.07, 0.04]	−0.08 [−0.12, −0.05]*	−0.13 [−0.17, −0.08]*	0.04 [0.02, 0.06]*
discrep	0.09 [0.02, 0.16]	−0.06 [−0.12, −0.01]	−0.09 [−0.13, −0.05]*	0.12 [0.07, 0.16]*	0.01 [−0.01, 0.03]
tentat	0.03 [−0.04, 0.09]	−0.00 [−0.06, 0.06]	0.06 [0.02, 0.10]	−0.01 [−0.06, 0.04]	0.01 [−0.02, 0.03]
certitude	−0.06 [−0.13, 0.00]	−0.01 [−0.06, 0.05]	0.05 [0.01, 0.09]	0.10 [0.05, 0.14]*	−0.04 [−0.06, −0.02]*
lexical_diversity	0.09 [0.02, 0.16]	−0.12 [−0.17, −0.06]*	−0.21 [−0.25, −0.17]*	−0.05 [−0.10, −0.00]	−0.02 [−0.05, 0.00]
difficult_words	−0.04 [−0.11, 0.03]	0.12 [0.06, 0.17]*	0.25 [0.21, 0.28]*	0.06 [0.01, 0.11]	0.02 [−0.00, 0.04]

B.3 Prompt Design

The detailed version of the zeroshot prompt included additional instructions about cues of deception. For fewshot prompts, the `<fewshot_examples>` tag is replaced with two examples each for deceptive and truthful instances, selected based on the sampling logic.

> **System:** *You are an expert deception detection analyst. Your task is to determine whether the author of a text is being intentionally deceptive. A statement is deceptive when the author deliberately expresses something they believe to be false or misrepresents their real thoughts, feelings, intentions, or actions. Focus on the author's intent rather than factual correctness. A person can be honest but mistaken if they sincerely believe what they say, and they can be deceptive while stating something true if they do not genuinely believe it. Use linguistic, emotional, and contextual indicators to infer intent.*
>
> **User:** *Evaluate whether the author of the provided text is being intentionally deceptive. Consider the author deceptive if they knowingly misrepresent their beliefs, emotions, or intentions. Analyze the following dimensions:*
> ***Intention over accuracy**: decide whether the author actually believes their statements.*
> ***Internal contradictions**: look for inconsistencies within the text or between claims and context.* ***Strategic framing**: detect signs of manipulation, concealment, persuasion, or image management.* ***Language indicators**: identify vagueness, distancing, over-justification, rehearsed tone, or unnatural certainty.* ***Emotional consistency**: check if tone and emotional content match the claimed feelings or stakes.* ***Contextual probability**: judge whether the described situation makes sense if the author is honest.* ***Ambiguity handling**: being wrong or exaggerating is not lying unless knowingly false or strategically misleading.*
> *If honesty and deception are both plausible, select the more likely interpretation and explain why. Possible scenarios: honest but mistaken about facts means not lying; factually true but not believed means lying; selective omission or framing to mislead means lying; pretending emotions or intentions means lying.*
> `<fewshot_examples>`
> *Provide your output strictly in the following JSON format use quotes consistently. Strictly use single quotes for the keys and values in the JSON as shown below:*
> *{ 'is_lying': true/false, 'reason_title': Very brief summary (as a short phrase) of why you believe the author is or isn't lying, 'explanation': Detailed explanation supporting your judgment }*
> *Text to evaluate:* `<input_instance>`

For chain-of-thought reasoning based prompts. The section after the `<few-shot_examples>` is replaced with the following section to encourage reasoning before predicting the label.

> *Think step by step: first analyse the text for deception cues and present your reasoning on whether the text is deceptive or not. Provide your output strictly in the following JSON format use quotes consistently. Strictly use single quotes for the keys and values in the JSON as shown below:*
> *{ 'explanation': detailed reasoning including cues, contradictions, and inferred intent, 'reason_title': 'very brief phrase explaining the main reason' 'is_lying': true/false, }*
> *Text to evaluate:* `<input_instance>`

B.4 Deception Detection Results

Tables B.5 and Table B.6 show the performance of classifiers across all datasets on the training data and Tables B.7 and Table B.8 show validation results on holdout data. We observe that for the comparison datasets **BERT** model performs consistently well on all three datasets, with **Mistral-7B** outperforming **BERT** in OPEN_DOMAIN dataset.

Table B.4
Dataset composition across training and holdout splits. Each row represents a dataset, with counts of total, truthful, and deceptive instances per split.

Dataset	Train Split			Holdout Split		
	Total	**Truthful**	**Deceptive**	**Total**	**Truthful**	**Deceptive**
DEFABEL_V1_DE	828	301	527	203	87	116
DEFABEL_V2_DE	776	388	388	192	96	96
DEFABEL_V2_EN	654	327	327	150	75	75
CROSS_CULT_DE	800	400	400	400	200	200
OPEN_DOMAIN	5734	2851	2883	1434	733	701
INTENTION	1312	604	708	328	179	149
DEC_OP	2000	1000	1000	500	250	250

Large language models (LLMs) `Mistral-7b`, `Llama3-8B`, and `Qwen-14B` exhibit variable results, often favoring high recall on truthful instances but struggling with deceptive cases. Table B.4 shows the label distribution in the data used for training and evaluating the models.

Tables B.9, B.10, B.11, B.12, and B.13 show the results for prompt and model size variation experiments.

Table B.5
Performance across DEFABEL datasets and classifiers on **train** split. Precision, recall, and F1 scores with standard errors are grouped by class label.

Source	Classifier	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
DEFABEL_EN	MajorityBaseline	0.50 ± 0.01	0.50	1.00	0.67 ± 0.01	0.00	0.00	0.00	0.33 ± 0.01
	RandomBaseline	0.49 ± 0.06	0.49	0.51	0.50 ± 0.05	0.49	0.47	0.48 ± 0.06	0.49 ± 0.06
	LogisticRegression	0.51 ± 0.05	0.51	0.51	0.51 ± 0.05	0.51	0.51	0.51 ± 0.06	0.51 ± 0.05
	SVC	0.50 ± 0.06	0.50	0.50	0.50 ± 0.05	0.50	0.50	0.50 ± 0.09	0.50 ± 0.06
	RandomForest	0.52 ± 0.05	0.52	0.53	0.52 ± 0.05	0.52	0.50	0.51 ± 0.06	0.52 ± 0.05
	GradientBoosting	0.51 ± 0.06	0.51	0.51	0.51 ± 0.07	0.51	0.51	0.51 ± 0.07	0.51 ± 0.06
	BERT	0.48 ± 0.01	0.42	0.45	0.38 ± 0.08	0.57	0.53	0.43 ± 0.07	0.40 ± 0.03
	Mistral-7b	0.52	0.55	0.23	0.33	0.51	0.81	0.63	0.48
	Llama3-8B	0.52	0.60	0.09	0.16	0.51	0.94	0.66	0.41
	Qwen-14B	0.50	0.64	0.02	0.04	0.50	0.99	0.67	0.35
	GPT-4o-mini	0.50	0.48	0.12	0.20	0.50	0.87	0.63	0.41
DEFABEL_v1_DE	MajorityBaseline	0.64 ± 0.00	0.64	1.00	0.78 ± 0.00	0.00	0.00	0.00	0.50 ± 0.00
	RandomBaseline	0.53 ± 0.07	0.67	0.51	0.58 ± 0.07	0.40	0.56	0.47 ± 0.08	0.54 ± 0.07
	LogisticRegression	0.62 ± 0.05	0.66	0.85	0.74 ± 0.04	0.46	0.23	0.30 ± 0.06	0.58 ± 0.05
	SVC	0.63 ± 0.03	0.64	0.95	0.77 ± 0.02	0.47	0.07	0.12 ± 0.05	0.53 ± 0.03
	RandomForest	0.61 ± 0.03	0.64	0.91	0.75 ± 0.03	0.38	0.10	0.16 ± 0.04	0.53 ± 0.02
	GradientBoosting	0.61 ± 0.05	0.65	0.84	0.73 ± 0.04	0.44	0.21	0.29 ± 0.07	0.57 ± 0.05
	BERT	0.64 ± 0.01	0.65	0.95	0.77 ± 0.01	0.32	0.08	0.13 ± 0.04	0.54 ± 0.02
	Mistral-7b	0.39	0.79	0.06	0.12	0.37	0.97	0.54	0.27
	Llama3-8B	0.39	0.85	0.04	0.08	0.37	0.99	0.54	0.25
	Qwen-14B	0.43	0.81	0.15	0.25	0.39	0.94	0.55	0.36
	GPT-4o-mini	0.50	0.81	0.27	0.40	0.41	0.89	0.56	0.46
DEFABEL_v2_DE	MajorityBaseline	0.50 ± 0.00	0.50	1.00	0.67 ± 0.00	0.00	0.00	0.00	0.33 ± 0.00
	RandomBaseline	0.48 ± 0.03	0.48	0.46	0.47 ± 0.03	0.49	0.51	0.50 ± 0.03	0.48 ± 0.03
	LogisticRegression	0.51 ± 0.07	0.51	0.49	0.50 ± 0.08	0.51	0.54	0.52 ± 0.09	0.51 ± 0.08
	SVC	0.54 ± 0.03	0.54	0.51	0.52 ± 0.03	0.54	0.57	0.55 ± 0.04	0.54 ± 0.03
	RandomForest	0.51 ± 0.05	0.51	0.55	0.53 ± 0.03	0.51	0.47	0.49 ± 0.09	0.51 ± 0.05
	GradientBoosting	0.50 ± 0.03	0.50	0.47	0.49 ± 0.03	0.50	0.53	0.51 ± 0.07	0.50 ± 0.03
	BERT	0.46 ± 0.01	0.40	0.41	0.35 ± 0.07	0.48	0.56	0.45 ± 0.06	0.39 ± 0.02
	Mistral-7b	0.51	0.61	0.08	0.14	0.51	0.95	0.66	0.40
	Llama3-8B	0.52	0.73	0.06	0.11	0.51	0.98	0.67	0.39
	Qwen-14B	0.52	0.71	0.07	0.14	0.51	0.97	0.67	0.40
	GPT-4o-mini	0.54	0.68	0.15	0.25	0.52	0.93	0.67	0.46

Table B.6
Performance across English datasets and classifiers on **train** split. Precision, recall, and F1 scores with standard errors are grouped by class label.

Source	Classifier	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.50 ± 0.05	0.50	0.48	0.49 ± 0.05	0.50	0.53	0.51 ± 0.05	0.50 ± 0.05
	LogisticRegression	0.65 ± 0.08	0.64	0.66	0.65 ± 0.08	0.65	0.64	0.64 ± 0.08	0.65 ± 0.08
	SVC	0.64 ± 0.07	0.63	0.68	0.65 ± 0.07	0.65	0.61	0.63 ± 0.08	0.64 ± 0.07
	RandomForest	0.66 ± 0.07	0.65	0.69	0.67 ± 0.07	0.67	0.62	0.64 ± 0.08	0.66 ± 0.07
	GradientBoosting	0.66 ± 0.06	0.66	0.68	0.67 ± 0.06	0.67	0.65	0.65 ± 0.07	0.66 ± 0.06
	BERT	0.70 ± 0.02	0.69	0.78	0.72 ± 0.02	0.74	0.63	0.67 ± 0.04	0.70 ± 0.02
	Mistral-7b	0.60	0.70	0.33	0.45	0.56	0.86	0.68	0.57
	Llama3-8B	0.53	0.64	0.12	0.21	0.51	0.93	0.66	0.43
	Qwen-14B	0.52	0.83	0.05	0.09	0.51	0.99	0.67	0.38
	GPT-4o-mini	0.53	0.93	0.07	0.12	0.52	0.99	0.68	0.40
DEC_OP	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.52 ± 0.04	0.52	0.52	0.52 ± 0.04	0.52	0.52	0.52 ± 0.04	0.52 ± 0.04
	LogisticRegression	0.73 ± 0.03	0.72	0.73	0.73 ± 0.04	0.73	0.72	0.72 ± 0.03	0.72 ± 0.03
	SVC	0.74 ± 0.02	0.74	0.75	0.74 ± 0.02	0.74	0.73	0.74 ± 0.03	0.74 ± 0.02
	RandomForest	0.73 ± 0.03	0.72	0.75	0.73 ± 0.03	0.74	0.70	0.72 ± 0.02	0.73 ± 0.03
	GradientBoosting	0.73 ± 0.03	0.73	0.74	0.73 ± 0.04	0.73	0.73	0.73 ± 0.03	0.73 ± 0.03
	BERT	0.83 ± 0.01	0.85	0.79	0.82 ± 0.01	0.81	0.86	0.83 ± 0.01	0.83 ± 0.01
	Mistral-7b	0.76	0.89	0.59	0.71	0.69	0.93	0.79	0.75
	Llama3-8B	0.59	0.85	0.23	0.36	0.55	0.96	0.70	0.53
	Qwen-14B	0.51	0.92	0.01	0.02	0.50	1.00	0.67	0.35
	GPT-4o-mini	0.55	0.92	0.10	0.19	0.53	0.99	0.69	0.44
INTENTION	MajorityBaseline	0.54 ± 0.00	0.54	1.00	0.70 ± 0.00	0.00	0.00	0.00	0.38 ± 0.00
	RandomBaseline	0.49 ± 0.04	0.53	0.41	0.46 ± 0.04	0.45	0.58	0.51 ± 0.04	0.48 ± 0.04
	LogisticRegression	0.63 ± 0.04	0.65	0.69	0.67 ± 0.04	0.61	0.56	0.58 ± 0.04	0.63 ± 0.04
	SVC	0.64 ± 0.04	0.65	0.73	0.69 ± 0.03	0.63	0.54	0.58 ± 0.06	0.64 ± 0.04
	RandomForest	0.63 ± 0.03	0.64	0.73	0.68 ± 0.03	0.62	0.51	0.56 ± 0.04	0.63 ± 0.03
	GradientBoosting	0.61 ± 0.04	0.63	0.68	0.65 ± 0.04	0.58	0.52	0.55 ± 0.04	0.61 ± 0.04
	BERT	0.70 ± 0.01	0.71	0.76	0.73 ± 0.01	0.70	0.64	0.66 ± 0.01	0.70 ± 0.01
	Mistral-7b	0.46	0.52	0.07	0.13	0.46	0.92	0.61	0.35
	Llama3-8B	0.46	0.53	0.01	0.02	0.46	0.99	0.63	0.30
	Qwen-14B	0.46	0.75	0.01	0.02	0.46	1.00	0.63	0.30
	GPT-4o-mini	0.46	1.00	0.00	0.00	0.46	1.00	0.63	0.29
OPEN_DOMAIN	MajorityBaseline	0.50 ± 0.00	0.50	1.00	0.67 ± 0.00	0.00	0.00	0.00	0.34 ± 0.00
	RandomBaseline	0.50 ± 0.03	0.50	0.48	0.49 ± 0.03	0.50	0.51	0.51 ± 0.03	0.50 ± 0.03
	LogisticRegression	0.58 ± 0.01	0.60	0.52	0.56 ± 0.02	0.57	0.65	0.61 ± 0.01	0.58 ± 0.01
	SVC	0.58 ± 0.01	0.58	0.59	0.59 ± 0.01	0.58	0.56	0.57 ± 0.01	0.58 ± 0.01
	RandomForest	0.55 ± 0.02	0.55	0.59	0.57 ± 0.02	0.56	0.52	0.53 ± 0.02	0.55 ± 0.02
	GradientBoosting	0.57 ± 0.01	0.57	0.58	0.57 ± 0.01	0.57	0.56	0.56 ± 0.02	0.57 ± 0.01
	BERT	0.67 ± 0.03	0.72	0.55	0.59 ± 0.10	0.65	0.79	0.70 ± 0.01	0.64 ± 0.06
	Mistral-7b	0.70	0.73	0.63	0.68	0.67	0.76	0.71	0.70
	Llama3-8B	0.59	0.87	0.22	0.35	0.55	0.97	0.70	0.53
	Qwen-14B	0.55	0.90	0.11	0.20	0.52	0.99	0.68	0.44
	GPT-4o-mini	0.63	0.94	0.29	0.44	0.58	0.98	0.73	0.58

Table B.7
Performance across DEFABEL datasets and classifiers on **holdout** split. Precision, recall, and F1 scores with standard errors are grouped by class label.

Source	Classifier	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
DEFABEL_EN	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.51	0.51	0.47	0.49	0.51	0.56	0.54	0.51
	LogisticRegression	0.50	0.50	0.64	0.56	0.50	0.36	0.42	0.49
	SVC	0.53	0.55	0.31	0.39	0.52	0.75	0.61	0.50
	RandomForest	0.47	0.48	0.56	0.52	0.47	0.39	0.42	0.47
	GradientBoosting	0.51	0.51	0.55	0.53	0.51	0.48	0.50	0.51
	BERT	0.53	0.52	0.69	0.59	0.54	0.36	0.43	0.51
	Mistral-7b	0.51	0.55	0.15	0.23	0.51	0.88	0.64	0.44
	Llama3-8B	0.52	0.80	0.05	0.10	0.51	0.99	0.67	0.39
	Qwen-14B	0.51	0.60	0.04	0.07	0.50	0.97	0.66	0.37
	GPT-4o-mini	0.47	0.37	0.09	0.15	0.48	0.84	0.61	0.38
DEFABEL_v1_DE	MajorityBaseline	0.57	0.57	1.00	0.73	0.00	0.00	0.00	0.36
	RandomBaseline	0.46	0.53	0.47	0.50	0.39	0.45	0.41	0.46
	LogisticRegression	0.58	0.59	0.85	0.70	0.51	0.21	0.30	0.50
	SVC	0.56	0.57	0.91	0.70	0.41	0.08	0.13	0.42
	RandomForest	0.56	0.57	0.94	0.71	0.42	0.06	0.10	0.41
	GradientBoosting	0.52	0.56	0.81	0.66	0.35	0.14	0.20	0.43
	BERT	0.57	0.57	1.00	0.73	0.00	0.00	0.00	0.42
	Mistral-7b	0.45	0.64	0.08	0.14	0.43	0.94	0.59	0.33
	Llama3-8B	0.43	0.60	0.03	0.05	0.43	0.98	0.60	0.28
	Qwen-14B	0.47	0.62	0.20	0.30	0.44	0.84	0.58	0.42
	GPT-4o-mini	0.51	0.70	0.26	0.38	0.46	0.85	0.60	0.47
DEFABEL_v2_DE	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.51	0.51	0.50	0.50	0.51	0.51	0.51	0.51
	LogisticRegression	0.53	0.53	0.54	0.53	0.53	0.51	0.52	0.53
	SVC	0.48	0.48	0.52	0.50	0.48	0.44	0.46	0.48
	RandomForest	0.46	0.46	0.55	0.50	0.45	0.36	0.40	0.45
	GradientBoosting	0.49	0.49	0.51	0.50	0.49	0.48	0.49	0.49
	BERT	0.55	0.56	0.46	0.51	0.54	0.65	0.59	0.55
	Mistral-7b	0.46	0.23	0.03	0.06	0.48	0.90	0.63	0.34
	Llama3-8B	0.48	0.33	0.03	0.06	0.49	0.94	0.65	0.35
	Qwen-14B	0.50	0.50	0.03	0.06	0.50	0.97	0.66	0.36
	GPT-4o-mini	0.55	0.74	0.15	0.24	0.53	0.95	0.68	0.46

Table B.8
Performance across English comparison datasets and classifiers on **holdout** split. Precision, recall, and F1 scores with standard errors are grouped by class label.

Source	Classifier	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.48	0.48	0.47	0.48	0.48	0.48	0.48	0.48
	LogisticRegression	0.59	0.58	0.68	0.62	0.61	0.50	0.55	0.59
	SVC	0.57	0.56	0.69	0.61	0.59	0.46	0.51	0.56
	RandomForest	0.65	0.62	0.79	0.69	0.71	0.52	0.60	0.65
	GradientBoosting	0.66	0.64	0.70	0.67	0.67	0.61	0.64	0.65
	BERT	0.62	0.58	0.83	0.69	0.71	0.40	0.51	0.60
	Mistral-7b	0.49	0.48	0.21	0.30	0.50	0.77	0.60	0.45
	Llama3-8B	0.49	0.25	0.01	0.02	0.49	0.97	0.66	0.34
	Qwen-14B	0.51	1.00	0.01	0.03	0.50	1.00	0.67	0.35
	GPT-4o-mini	0.51	1.00	0.01	0.03	0.50	1.00	0.67	0.35
DEC_OP	MajorityBaseline	0.50	0.50	1.00	0.67	0.00	0.00	0.00	0.33
	RandomBaseline	0.53	0.53	0.52	0.53	0.53	0.54	0.54	0.53
	LogisticRegression	0.74	0.74	0.74	0.74	0.74	0.74	0.74	0.74
	SVC	0.71	0.70	0.74	0.72	0.73	0.68	0.71	0.71
	RandomForest	0.73	0.72	0.75	0.74	0.74	0.71	0.72	0.73
	GradientBoosting	0.72	0.74	0.70	0.72	0.71	0.75	0.73	0.72
	BERT	0.78	0.92	0.61	0.73	0.71	0.95	0.81	0.77
	Mistral-7b	0.62	0.91	0.28	0.42	0.57	0.97	0.72	0.57
	Llama3-8B	0.55	0.93	0.10	0.18	0.52	0.99	0.69	0.43
	Qwen-14B	0.50	0.50	0.00	0.01	0.50	1.00	0.67	0.34
	GPT-4o-mini	0.51	0.88	0.03	0.05	0.51	1.00	0.67	0.36
INTENTION	MajorityBaseline	0.45	0.45	1.00	0.62	0.00	0.00	0.00	0.31
	RandomBaseline	0.45	0.41	0.45	0.43	0.50	0.46	0.48	0.45
	LogisticRegression	0.62	0.57	0.66	0.61	0.68	0.59	0.63	0.62
	SVC	0.62	0.56	0.70	0.62	0.69	0.54	0.61	0.62
	RandomForest	0.63	0.57	0.77	0.66	0.73	0.52	0.61	0.63
	GradientBoosting	0.62	0.57	0.70	0.63	0.69	0.55	0.61	0.62
	BERT	0.70	0.65	0.70	0.68	0.74	0.69	0.71	0.70
	Mistral-7b	0.57	0.68	0.09	0.15	0.56	0.97	0.71	0.46
	Llama3-8B	0.55	0.50	0.01	0.01	0.55	0.99	0.70	0.39
	Qwen-14B	0.54	0.00	0.00	0.00	0.54	0.99	0.70	0.38
	GPT-4o-mini	0.45	0.45	1.00	0.62	0.00	0.00	0.00	0.28
OPEN_DOMAIN	MajorityBaseline	0.49	0.49	1.00	0.66	0.00	0.00	0.00	0.33
	RandomBaseline	0.49	0.48	0.49	0.48	0.50	0.49	0.49	0.49
	LogisticRegression	0.58	0.58	0.52	0.55	0.58	0.63	0.61	0.58
	SVC	0.58	0.56	0.58	0.57	0.59	0.57	0.58	0.58
	RandomForest	0.58	0.57	0.62	0.59	0.60	0.55	0.57	0.58
	GradientBoosting	0.58	0.57	0.57	0.57	0.59	0.58	0.59	0.58
	BERT	0.72	0.73	0.66	0.70	0.70	0.77	0.74	0.72
	Mistral-7b	0.71	0.73	0.65	0.69	0.70	0.77	0.73	0.71
	Llama3-8B	0.62	0.92	0.25	0.39	0.58	0.98	0.73	0.56
	Qwen-14B	0.56	0.89	0.11	0.19	0.54	0.99	0.70	0.45
	GPT-4o-mini	0.63	0.93	0.27	0.42	0.59	0.98	0.73	0.58

Table B.9
Evaluation metrics for **GPT-4o-mini**: accuracy, class-wise precision/recall/F1, and overall F1. **Bold** indicates the best value within each source across prompt types.

Source	Prompt type	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	zero_shot_simple	.507	**1.000**	.015	.030	.504	**1.000**	**.670**	.350
	zero_shot_detailed	.522	.558	.215	.310	.514	.830	.635	.473
	fewshot_random	**.542**	.655	.180	.282	**.525**	.905	.664	.473
	fewshot_feature_similarity	.515	.537	.220	.312	.509	.810	.625	.469
	fewshot_semantic_similarity	.520	.561	.185	.278	.512	.855	.640	.459
	zero_shot_detailed_COT	.435	.432	.415	.423	.438	.455	.446	.435
	fewshot_random_COT	.468	.464	.415	.438	.471	.520	.494	.466
	fewshot_feature_similarity_COT	.410	.409	.405	.407	.411	.415	.413	.410
	fewshot_semantic_similarity_COT	.475	.473	**.440**	**.456**	.477	.510	.493	**.474**
DEC_OP	zero_shot_simple	.512	.875	.028	.054	.506	**.996**	.671	.363
	zero_shot_detailed	.664	.836	.408	.548	.608	.920	.732	.640
	fewshot_random	.674	**.907**	.388	.543	.611	.960	.747	.645
	fewshot_feature_similarity	.668	.850	.408	.551	.611	.928	.737	.644
	fewshot_semantic_similarity	**.756**	.852	.620	**.718**	**.701**	.892	**.785**	**.751**
	zero_shot_detailed_COT	.668	.708	.572	.633	.641	.764	.697	.665
	fewshot_random_COT	.674	.792	.472	.591	.624	.876	.729	.660
	fewshot_feature_similarity_COT	.688	.740	.580	.650	.655	.796	.718	.684
	fewshot_semantic_similarity_COT	.694	.724	**.628**	.672	.671	.760	.713	.693
DEFABEL_EN	zero_shot_simple	.467	.368	.093	.149	.481	**.840**	**.612**	.380
	zero_shot_detailed	.467	.436	.227	.298	.477	.707	.570	.434
	fewshot_random	.453	.400	.187	.255	.470	.720	.568	.411
	fewshot_feature_similarity	.480	.462	.240	.316	.486	.720	.581	.448
	fewshot_semantic_similarity	.487	.480	.320	.384	.490	.653	.560	.472
	zero_shot_detailed_COT	.493	.492	.413	.449	.494	.573	.531	.490
	fewshot_random_COT	**.533**	**.535**	**.507**	**.521**	**.532**	.560	.545	**.533**
	fewshot_feature_similarity_COT	.527	.531	.453	.489	.523	.600	.559	.524
	fewshot_semantic_similarity_COT	.493	.494	**.507**	.500	.493	.480	.486	.493
DEFABEL_v1_DE	zero_shot_simple	.512	.698	.259	.377	.463	.851	.599	.488
	zero_shot_detailed	.483	.634	.224	.331	.444	.828	.578	.455
	fewshot_random	.488	.560	.483	.519	.417	.494	.453	.486
	fewshot_feature_similarity	**.527**	**.727**	.276	.400	**.472**	**.862**	**.610**	.505
	fewshot_semantic_similarity	**.527**	.591	.560	.575	.452	.483	.467	**.521**
	zero_shot_detailed_COT	.463	.541	.397	.458	.407	.552	.468	.463
	fewshot_random_COT	**.527**	.572	**.681**	**.622**	.431	.322	.368	.495
	fewshot_feature_similarity_COT	.502	.584	.448	.507	.439	.575	.498	.502
	fewshot_semantic_similarity_COT	.478	.535	.664	.592	.339	.230	.274	.433
DEFABEL_v2_DE	zero_shot_simple	.547	**.737**	.146	.243	.526	**.948**	**.677**	.460
	zero_shot_detailed	.531	.600	.188	.286	.519	.875	.651	.468
	fewshot_random	.531	.588	.208	.308	.519	.854	.646	.477
	fewshot_feature_similarity	.526	.586	.177	.272	.515	.875	.649	.460
	fewshot_semantic_similarity	.542	.553	.438	.488	.534	.646	.585	.537
	zero_shot_detailed_COT	.516	.519	.417	.462	.513	.615	.559	.511
	fewshot_random_COT	.552	.550	.573	.561	.554	.531	.543	.552
	fewshot_feature_similarity_COT	.510	.514	.396	.447	.508	.625	.561	.504
	fewshot_semantic_similarity_COT	**.573**	.569	**.604**	**.586**	**.578**	.542	.559	**.572**
INTENTION	zero_shot_simple	.546	.000	.000	.000	.546	**1.000**	.706	**.706**
	zero_shot_detailed	.552	**1.000**	.013	.026	.549	**1.000**	.709	.368
	fewshot_random	**.561**	**1.000**	.034	.065	.554	**1.000**	**.713**	.389
	fewshot_feature_similarity	**.561**	**1.000**	.034	.065	.554	**1.000**	**.713**	.389
	fewshot_semantic_similarity	.558	.667	.054	.099	.554	.978	.707	.403
	zero_shot_detailed_COT	.543	.474	.060	.107	.547	.944	.693	.400
	fewshot_random_COT	.549	.522	.081	.140	.551	.939	.694	.417
	fewshot_feature_similarity_COT	.558	.700	.047	.088	.553	.983	.708	.398
	fewshot_semantic_similarity_COT	**.561**	.576	**.128**	**.209**	**.559**	.922	.696	.452
OPEN_DOMAIN	zero_shot_simple	.635	.932	.272	.422	.585	**.981**	.733	.577
	zero_shot_detailed	.653	**.932**	.312	.468	.598	.978	.742	.605
	fewshot_random	.679	.908	.381	.537	.619	.963	.754	.645
	fewshot_feature_similarity	.658	.930	.324	.480	.602	.977	.745	.613
	fewshot_semantic_similarity	**.728**	.833	**.555**	**.666**	**.677**	.894	**.771**	**.718**
	zero_shot_detailed_COT	.649	.887	.324	.474	.598	.960	.737	.606
	fewshot_random_COT	.663	.854	.375	.521	.611	.939	.740	.631
	fewshot_feature_similarity_COT	.644	.880	.314	.463	.594	.959	.733	.598
	fewshot_semantic_similarity_COT	.700	.861	.461	.600	.643	.929	.760	.680

Table B.10
Evaluation metrics for **LLaMA3-8B**: accuracy, class-wise precision/recall/F1, and overall F1. **Bold** indicates the best value within each source across prompt types.

Source	Prompt type	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	zero_shot_simple	.490	.250	.010	.019	.495	**.970**	**.655**	.337
	zero_shot_detailed	.468	.397	.125	.190	.481	.810	.603	.397
	fewshot_random	.512	.514	.465	.488	.511	.560	.535	.511
	fewshot_feature_similarity	.468	.397	.125	.190	.481	.810	.603	.397
	fewshot_semantic_similarity	**.527**	**.536**	.410	.465	**.522**	.645	.577	**.521**
	zero_shot_detailed_COT	.475	.483	.700	.571	.455	.250	.323	.447
	fewshot_random_COT	.497	.498	**.830**	**.623**	.493	.165	.247	.435
	fewshot_feature_similarity_COT	.475	.483	.700	.571	.455	.250	.323	.447
	fewshot_semantic_similarity_COT	.507	.505	.750	.604	.515	.265	.350	.477
DEC_OP	zero_shot_simple	.546	**.926**	.100	.181	.524	**.992**	.686	.433
	zero_shot_detailed	.614	.835	.284	.424	.569	.944	.710	.567
	fewshot_random	**.800**	.899	.676	**.772**	.740	.924	**.822**	**.797**
	fewshot_feature_similarity	.614	.835	.284	.424	.569	.944	.710	.567
	fewshot_semantic_similarity	.742	.834	.604	.701	.690	.880	.773	.737
	zero_shot_detailed_COT	.646	.616	.776	.687	.697	.516	.593	.640
	fewshot_random_COT	.672	.634	.812	.712	.739	.532	.619	.665
	fewshot_feature_similarity_COT	.646	.616	.776	.687	.697	.516	.593	.640
	fewshot_semantic_similarity_COT	.672	.629	**.840**	.719	**.759**	.504	.606	.662
DEFABEL_EN	zero_shot_simple	.520	**.800**	.053	.100	.510	**.987**	**.673**	.386
	zero_shot_detailed	.520	.560	.187	.280	.512	.853	.640	.460
	fewshot_random	.500	.500	.400	.444	.500	.600	.545	.495
	fewshot_feature_similarity	.520	.560	.187	.280	.512	.853	.640	.460
	fewshot_semantic_similarity	.493	.492	.427	.457	.494	.560	.525	.491
	zero_shot_detailed_COT	.513	.512	.573	.541	.515	.453	.482	.512
	fewshot_random_COT	.507	.505	.627	.560	.509	.387	.439	.499
	fewshot_feature_similarity_COT	.513	.512	.573	.541	.515	.453	.482	.512
	fewshot_semantic_similarity_COT	**.547**	.535	**.707**	**.609**	**.569**	.387	.460	**.535**
DEFABEL_v1_DE	zero_shot_simple	.433	**.600**	.026	.050	.429	**.977**	**.596**	.323
	zero_shot_detailed	.453	.581	.155	.245	.430	.851	.571	.408
	fewshot_random	.542	.597	.612	.604	.464	.448	.456	**.530**
	fewshot_feature_similarity	.453	.581	.155	.245	.430	.851	.571	.408
	fewshot_semantic_similarity	.552	.593	.690	.637	.471	.368	.413	.525
	zero_shot_detailed_COT	.480	.550	.478	.512	.412	.483	.444	.478
	fewshot_random_COT	**.562**	.581	**.836**	**.686**	**.472**	.195	.276	.481
	fewshot_feature_similarity_COT	.480	.550	.478	.512	.412	.483	.444	.478
	fewshot_semantic_similarity_COT	.547	.571	.828	.676	.429	.172	.246	.461
DEFABEL_v2_DE	zero_shot_simple	.484	.333	.031	.057	.492	**.938**	**.645**	.351
	zero_shot_detailed	.490	.464	.135	.210	.494	.844	.623	.416
	fewshot_random	.547	.545	.573	.558	.549	.521	.535	.547
	fewshot_feature_similarity	.490	.464	.135	.210	.494	.844	.623	.416
	fewshot_semantic_similarity	**.557**	**.558**	.552	.555	**.557**	.562	.560	**.557**
	zero_shot_detailed_COT	.474	.477	.542	.507	.470	.406	.436	.472
	fewshot_random_COT	.484	.488	.615	.544	.479	.354	.407	.475
	fewshot_feature_similarity_COT	.474	.477	.542	.507	.470	.406	.436	.472
	fewshot_semantic_similarity_COT	.516	.510	**.771**	**.614**	.532	.260	.350	.482
INTENTION	zero_shot_simple	.546	.500	.007	.013	.546	**.994**	.705	.359
	zero_shot_detailed	.558	**.750**	.040	.076	.553	.989	**.709**	.393
	fewshot_random	.558	.559	.128	.208	.558	.916	.693	.451
	fewshot_feature_similarity	.558	**.750**	.040	.076	.553	.989	**.709**	.393
	fewshot_semantic_similarity	**.585**	.597	.268	.370	.582	.849	.691	.531
	zero_shot_detailed_COT	.546	.500	.416	.454	.574	.654	.611	.533
	fewshot_random_COT	.540	.494	.517	.505	.581	.559	.570	**.537**
	fewshot_feature_similarity_COT	.546	.500	.416	.454	.574	.654	.611	.533
	fewshot_semantic_similarity_COT	.521	.481	**.685**	**.565**	**.595**	.385	.468	.516
OPEN_DOMAIN	zero_shot_simple	.621	**.920**	.245	.387	.576	**.980**	.725	.556
	zero_shot_detailed	.654	.823	.372	.513	.606	.924	.732	.622
	fewshot_random	.697	.801	.505	.619	.650	.880	.748	.684
	fewshot_feature_similarity	.654	.823	.372	.513	.606	.924	.732	.622
	fewshot_semantic_similarity	**.707**	.817	.516	.633	.658	.889	**.756**	.695
	zero_shot_detailed_COT	.673	.760	.481	.589	.634	.855	.728	.658
	fewshot_random_COT	.702	.772	.553	.644	.665	.844	.744	.694
	fewshot_feature_similarity_COT	.673	.760	.481	.589	.634	.855	.728	.658
	fewshot_semantic_similarity_COT	.706	.753	**.595**	**.665**	**.677**	.813	.739	**.702**

Table B.11
Evaluation metrics for **Mistral-7B**: accuracy, class-wise precision/recall/F1, and overall F1. **Bold** indicates the best value within each source across prompt types.

Source	Prompt type	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	zero_shot_simple	.492	.483	.215	.298	.495	.770	.603	.450
	zero_shot_detailed	.504	.508	.348	.413	.502	.661	.571	**.492**
	fewshot_random	.519	.530	.256	.345	.515	.777	.620	.482
	fewshot_feature_similarity	.504	.508	.348	.413	.502	.661	.571	**.492**
	fewshot_semantic_similarity	**.531**	**.549**	.224	.318	.526	**.824**	**.642**	.480
	zero_shot_detailed_COT	.492	.497	**.932**	.648	.409	.048	.085	.367
	fewshot_random_COT	.524	.512	.813	.628	.568	.241	.338	.483
	fewshot_feature_similarity_COT	.492	.497	**.932**	.648	.409	.048	.085	.367
	fewshot_semantic_similarity_COT	.522	.516	.922	**.662**	**.583**	.112	.188	.425
DEC_OP	zero_shot_simple	.624	**.908**	.276	.423	.573	**.972**	.721	.572
	zero_shot_detailed	**.699**	.844	.505	.632	.635	.903	**.746**	**.689**
	fewshot_random	.692	.876	.474	.615	.621	.928	.744	.679
	fewshot_feature_similarity	**.699**	.844	.505	.632	.635	.903	**.746**	**.689**
	fewshot_semantic_similarity	.640	.816	.447	.578	.564	.876	.687	.632
	zero_shot_detailed_COT	.528	.519	**.983**	.679	**.765**	.057	.107	.393
	fewshot_random_COT	.546	.538	.909	.676	.607	.152	.244	.460
	fewshot_feature_similarity_COT	.528	.519	**.983**	.679	**.765**	.057	.107	.393
	fewshot_semantic_similarity_COT	.551	.536	.927	**.680**	.667	.155	.251	.465
DEFABEL_EN	zero_shot_simple	.513	**.550**	.147	.232	.508	**.880**	**.644**	.438
	zero_shot_detailed	.476	.412	.096	.156	.484	.861	.620	.388
	fewshot_random	**.517**	.541	.278	.367	.509	.761	.610	**.489**
	fewshot_feature_similarity	.476	.412	.096	.156	.484	.861	.620	.388
	fewshot_semantic_similarity	.504	.500	.243	.327	.505	.761	.607	.467
	zero_shot_detailed_COT	.472	.480	.857	.615	.412	.097	.157	.386
	fewshot_random_COT	.490	.500	.904	.644	.364	.057	.099	.371
	fewshot_feature_similarity_COT	.472	.480	.857	.615	.412	.097	.157	.386
	fewshot_semantic_similarity_COT	.510	.507	**.972**	**.667**	**.600**	.042	.079	.373
DEFABEL_v1_DE	zero_shot_simple	.448	**.643**	.078	.138	.434	**.943**	**.594**	.366
	zero_shot_detailed	.470	.606	.179	.276	.442	.849	.582	.429
	fewshot_random	.472	.560	.372	.447	.418	.607	.495	.471
	fewshot_feature_similarity	.470	.606	.179	.276	.442	.849	.582	.429
	fewshot_semantic_similarity	.480	.554	.411	.472	.426	.570	.488	.480
	zero_shot_detailed_COT	**.576**	.578	.974	.725	**.500**	.036	.067	.396
	fewshot_random_COT	.567	.569	.991	.723	.000	.000	.000	.723
	fewshot_feature_similarity_COT	**.576**	.578	.974	.725	**.500**	.036	.067	.396
	fewshot_semantic_similarity_COT	.574	.574	**1.000**	**.730**	.000	.000	.000	**.730**
DEFABEL_v2_DE	zero_shot_simple	.464	.231	.031	.055	.480	**.896**	.625	.340
	zero_shot_detailed	.516	.500	.112	.183	.518	.895	.656	.420
	fewshot_random	**.533**	**.556**	.169	.259	.529	.874	**.659**	.459
	fewshot_feature_similarity	.516	.500	.112	.183	.518	.895	.656	.420
	fewshot_semantic_similarity	.514	.508	.344	.411	.516	.677	.586	**.498**
	zero_shot_detailed_COT	.494	.494	.966	.654	.500	.033	.062	.358
	fewshot_random_COT	.508	.506	.957	.662	.556	.054	.099	.380
	fewshot_feature_similarity_COT	.494	.494	.966	.654	.500	.033	.062	.358
	fewshot_semantic_similarity_COT	.505	.503	**.979**	**.664**	**.600**	.032	.060	.362
INTENTION	zero_shot_simple	.567	**.684**	.087	.155	.560	**.966**	**.709**	.432
	zero_shot_detailed	.533	.273	.025	.045	.544	.946	.691	.368
	fewshot_random	.510	.412	.061	.107	.518	.920	.663	.385
	fewshot_feature_similarity	.533	.273	.025	.045	.544	.946	.691	.368
	fewshot_semantic_similarity	**.590**	.603	.346	.440	.585	.801	.676	**.558**
	zero_shot_detailed_COT	.462	.451	.914	.604	.571	.093	.160	.382
	fewshot_random_COT	.483	.474	.960	**.635**	**.643**	.064	.116	.375
	fewshot_feature_similarity_COT	.462	.451	.914	.604	.571	.093	.160	.382
	fewshot_semantic_similarity_COT	.440	.444	**.962**	.608	.286	.012	.024	.316
OPEN_DOMAIN	zero_shot_simple	.711	.727	.648	.686	.698	.770	.732	.709
	zero_shot_detailed	.719	**.874**	.454	.598	.670	**.944**	.784	.691
	fewshot_random	**.735**	.825	.545	.657	.695	.899	**.784**	**.720**
	fewshot_feature_similarity	.719	**.874**	.454	.598	.670	**.944**	.784	.691
	fewshot_semantic_similarity	.712	.758	.587	.662	.685	.828	.750	.706
	zero_shot_detailed_COT	.636	.610	.812	.696	.694	.451	.547	.622
	fewshot_random_COT	.615	.579	.914	.709	.768	.299	.431	.570
	fewshot_feature_similarity_COT	.636	.610	.812	.696	.694	.451	.547	.622
	fewshot_semantic_similarity_COT	.607	.567	**.947**	**.709**	**.829**	.261	.396	.553

Table B.12
Evaluation metrics for GPT-4o: accuracy, class-wise precision/recall/F1, and overall F1.
Bold indicates the best value within each source across prompt types.

Source	Prompt type	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	zero_shot_detailed	.497	.467	.035	.065	.499	.960	.656	.361
	fewshot_random	.505	.529	**.090**	**.154**	.503	.920	.650	**.402**
	fewshot_semantic_similarity	**.517**	**.684**	.065	.119	**.509**	**.970**	**.668**	.393
	zero_shot_detailed_COT	.512	.609	.070	.126	.507	.955	.662	.394
	fewshot_random_COT	.505	.538	.070	.124	.503	.940	.655	.389
	fewshot_semantic_similarity_COT	.510	.625	.050	.093	.505	**.970**	.664	.378
DEC_OP	zero_shot_detailed	.548	**.929**	.104	.187	.525	**.992**	.687	.437
	fewshot_random	.576	.913	.168	.284	.542	.984	.699	.491
	fewshot_semantic_similarity	.582	.902	.184	.306	.546	.980	.701	.503
	zero_shot_detailed_COT	.552	.810	.136	.233	.528	.968	.684	.458
	fewshot_random_COT	.568	.854	.164	.275	.538	.972	.692	.484
	fewshot_semantic_similarity_COT	**.584**	.889	**.192**	**.316**	**.547**	.976	**.701**	**.508**
DEFABEL_EN	zero_shot_detailed	.507	.529	.120	.196	.504	.893	.644	.420
	fewshot_random	**.553**	.643	**.240**	**.350**	**.533**	.867	.660	**.505**
	fewshot_semantic_similarity	.527	.562	**.240**	.336	.517	.813	.632	.484
	zero_shot_detailed_COT	.547	**.684**	.173	.277	.527	**.920**	**.670**	.473
	fewshot_random_COT	.500	.500	.147	.227	.500	.853	.631	.429
	fewshot_semantic_similarity_COT	.527	.577	.200	.297	.516	.853	.643	.470
DEFABEL_v1_DE	zero_shot_detailed	.478	.583	.302	.398	.434	.713	.539	.468
	fewshot_random	**.547**	.615	**.552**	**.582**	**.475**	.540	.505	**.544**
	fewshot_semantic_similarity	.527	.606	.491	.543	.459	.575	.510	.527
	zero_shot_detailed_COT	.458	.583	.181	.276	.431	**.828**	**.567**	.422
	fewshot_random_COT	.527	.614	.466	.529	.461	.609	.525	.527
	fewshot_semantic_similarity_COT	.522	**.630**	.397	.487	.462	.690	.553	.520
DEFABEL_v2_DE	zero_shot_detailed	.536	.667	.146	.239	.520	.927	.667	.453
	fewshot_random	.526	.619	.135	.222	.515	.917	.659	.441
	fewshot_semantic_similarity	**.578**	.642	**.354**	**.456**	**.554**	.802	.655	**.556**
	zero_shot_detailed_COT	.531	**.688**	.115	.196	.517	**.948**	**.669**	.433
	fewshot_random_COT	.516	.600	.094	.162	.508	.938	.659	.411
	fewshot_semantic_similarity_COT	.542	.643	.188	.290	.524	.896	.662	.476
INTENTION	zero_shot_detailed	.555	.615	.054	.099	.552	.972	.704	.402
	fewshot_random	.555	.714	.034	.064	.551	.989	.708	.386
	fewshot_semantic_similarity	**.591**	**.778**	**.141**	**.239**	**.575**	.966	**.721**	**.480**
	zero_shot_detailed_COT	.552	.667	.027	.052	.550	.989	.707	.379
	fewshot_random_COT	.552	.750	.020	.039	.549	**.994**	.708	.373
	fewshot_semantic_similarity_COT	.576	.708	.114	.197	.566	.961	.712	.454
OPEN_DOMAIN	zero_shot_detailed	.631	.881	.284	.429	.584	.963	.727	.578
	fewshot_random	.659	.863	.359	.508	.607	.945	.739	.623
	fewshot_semantic_similarity	**.713**	.849	**.504**	**.632**	**.658**	.914	**.765**	**.699**
	zero_shot_detailed_COT	.598	.937	.190	.316	.560	**.988**	.715	.515
	fewshot_random_COT	.623	**.944**	.243	.386	.577	.986	.728	.557
	fewshot_semantic_similarity_COT	.660	.884	.350	.501	.606	.956	.742	.621

Table B.13
Evaluation metrics for LLaMA3-70B: accuracy, class-wise precision/recall/F1, and overall F1. **Bold** indicates the best value within each source across prompt types.

Source	Prompt type	Accuracy	Deceptive			Truthful			F1 Overall
			Prec.	Recall	F1	Prec.	Recall	F1	
CROSS_CULT_DE	zero_shot_detailed	.507	**.714**	.025	.048	.504	**.990**	**.668**	.358
	fewshot_random	**.530**	.630	.145	.236	**.517**	.915	.661	**.448**
	fewshot_semantic_similarity	.522	.645	.100	.173	.512	.945	.664	.419
	zero_shot_detailed_COT	.507	.565	.065	.117	.504	.950	.659	.388
	fewshot_random_COT	.490	.476	**.195**	**.277**	.494	.785	.606	.441
	fewshot_semantic_similarity_COT	.519	.566	.152	.240	.512	.884	.648	.444
DEC_OP	zero_shot_detailed	.528	.889	.064	.119	.515	**.992**	.678	.398
	fewshot_random	.570	.872	.164	.276	.539	.976	.694	.485
	fewshot_semantic_similarity	.628	.864	.304	.450	.578	.952	.719	.584
	zero_shot_detailed_COT	.560	.895	.136	.236	.532	.984	.691	.464
	fewshot_random_COT	.598	**.930**	.212	.345	.555	.984	.710	.528
	fewshot_semantic_similarity_COT	**.680**	.862	**.427**	**.571**	**.620**	.932	**.745**	**.658**
DEFABEL_EN	zero_shot_detailed	.520	.600	.120	.200	.511	**.920**	**.657**	.429
	fewshot_random	.513	.529	.240	.330	.509	.787	.618	.474
	fewshot_semantic_similarity	.540	**.607**	.227	.330	.525	.853	.650	.490
	zero_shot_detailed_COT	.500	.500	.160	.242	.500	.840	.627	.435
	fewshot_random_COT	.520	.541	.267	.357	.513	.773	.617	.487
	fewshot_semantic_similarity_COT	**.553**	.595	**.333**	**.427**	**.537**	.773	.634	**.531**
DEFABEL_v1_DE	zero_shot_detailed	.463	.600	.181	.278	.435	.839	.573	.425
	fewshot_random	.493	.592	.362	.449	.439	.667	.530	.489
	fewshot_semantic_similarity	.488	.573	.405	.475	.430	.598	.500	.487
	zero_shot_detailed_COT	.470	.618	.183	.282	.440	**.851**	**.580**	.431
	fewshot_random_COT	.483	.566	.405	.472	.425	.586	.493	.483
	fewshot_semantic_similarity_COT	**.527**	**.625**	**.431**	**.510**	**.463**	.655	.543	**.527**
DEFABEL_v2_DE	zero_shot_detailed	.505	.556	.052	.095	.503	**.958**	**.659**	.377
	fewshot_random	**.526**	**.609**	.146	.235	.515	.906	.657	.446
	fewshot_semantic_similarity	.521	.562	.188	.281	.512	.854	.641	.461
	zero_shot_detailed_COT	.495	.444	.042	.076	.497	.948	.652	.364
	fewshot_random_COT	**.526**	**.609**	.146	.235	.515	.906	.657	.446
	fewshot_semantic_similarity_COT	**.526**	.551	**.281**	**.372**	**.517**	.771	.619	**.496**
INTENTION	zero_shot_detailed	.555	**1.000**	.020	.039	.551	**1.000**	.710	.375
	fewshot_random	.564	.750	.060	.112	.557	.983	.711	.411
	fewshot_semantic_similarity	**.616**	.767	**.221**	**.344**	**.593**	.944	**.728**	**.536**
	zero_shot_detailed_COT	.552	**1.000**	.013	.026	.549	**1.000**	.709	.368
	fewshot_random_COT	.558	.833	.034	.065	.553	.994	.711	.388
	fewshot_semantic_similarity_COT	.583	.731	.128	.218	.570	.961	.715	.467
OPEN_DOMAIN	zero_shot_detailed	.694	.920	.409	.566	.631	.966	.763	.665
	fewshot_random	.734	.902	.511	.653	.670	.947	.785	.719
	fewshot_semantic_similarity	**.757**	.865	**.596**	**.706**	**.702**	.911	**.793**	**.750**
	zero_shot_detailed_COT	.623	**.921**	.250	.393	.577	**.980**	.726	.560
	fewshot_random_COT	.671	.919	.358	.515	.612	.970	.751	.633
	fewshot_semantic_similarity_COT	.712	.898	.464	.611	.649	.949	.771	.691

Table B.14
Counts by error type (FP/FN), reasoning cluster, and dataset.

	Dataset	DEFABEL_v1_DE	DEFABEL_v2_DE	DEFABEL_EN	DEC_OP	OPEN_DOMAIN	CROSS_CULT_DE	INTENTION
Error Type	Reason Cluster Name							
FN	Contextual Misrepresentation	0	0	3	0	22	0	2
	Contradictions and Logical Inconsistencies	0	2	7	9	236	20	40
	Exaggeration or Oversimplification	0	1	4	12	21	18	0
	Generalization	0	0	0	2	19	0	0
	Lack of Evidence or Basis	0	0	0	0	49	0	0
	Misattribution or Attribution Errors	0	0	0	0	11	0	0
	Misinformation	5	4	14	1	43	0	6
	Misrepresentation of Facts or Science	9	22	44	56	428	63	6
	Outdated or Deprecated Information	0	0	0	0	22	0	0
FP	Common Beliefs and Scientific facts	11	10	22	47	30	11	2
	Consistency and Coherence	1	7	2	29	2	76	148
	Expressing Emotions	0	12	5	117	48	74	8
	Factually accurate claims	13	21	22	3	183	3	3
	Honest Tone with No Deception	221	128	54	112	556	85	572
	Hyperbole or Exaggeration	9	10	13	41	52	22	4
	Lack of Context	1	2	2	0	102	1	0
	Lack of Contradiction	5	3	4	6	26	34	11
	Misunderstanding or Misrepresentation	87	88	69	6	54	4	10
	Opinion and Subjectivity	0	2	0	1	80	0	1
	Personal Belief and Values	37	46	24	217	23	103	10
	Statements or Clarification	31	41	30	10	124	8	20
	Truthful Intent	185	81	68	4	0	3	3

B.5 Misclassification Analysis

Table B.14 presents the false positive (FP) and false negative (FN) counts for each rationale category (cluster) generated from the reasoning provided by the Mistral model, shown across the different datasets.

B.6 Cross-corpus Analysis

In Figure B.1, we visualize a cross-corpus comparison based on BERT-based deception classifiers. It shows the representations learned by the base model and multiple fine-tuned RoBERTa models using t-SNE projections of sentence embeddings. Each subplot shows the separation between deceptive and truthful texts across the English datasets, enabling comparison of how well different models capture deception-related distinctions. Rows correspond to models (base and fine-tuned), while columns represent the datasets being evaluated.

We observe that, along the diagonal, where each dataset is evaluated on the model that is fine-tuned with the same datasets, the separation between deceptive and truthful texts is generally more distinct in the case of most datasets except for DEFABEL_V2_EN. This visualized the model's difficulty to discern truthful and deceptive instances in belief-based datasets.

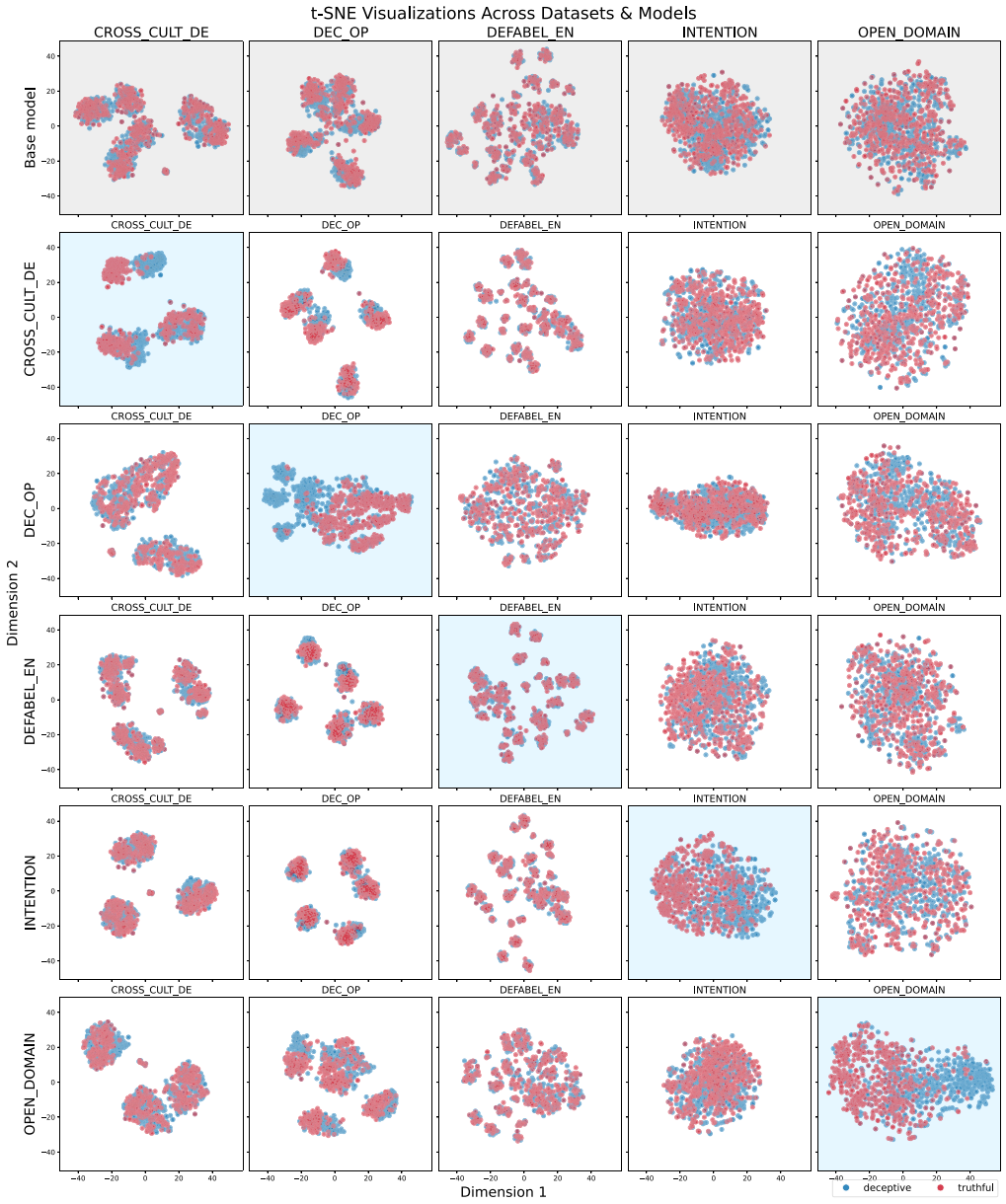


Figure B.1
t-SNE.

Acknowledgments

We gratefully acknowledge the insightful feedback provided by researchers at the Leibniz Institute for Psychology (ZPID), Trier, particularly with regard to our study design from a psychology perspective. We also extend our sincere thanks to Amelie Wührl (IT University of Copenhagen) for her valuable contributions to our discussions on the complex interplay between factuality and deception.

References

Almela, Ángela, Rafael Valencia-García, and Pascual Cantos. 2012. Seeing through deception: A computational approach to deceit detection in written communication. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 15–22.

Almela, Ángela. 2021. A corpus-based study of linguistic deception in Spanish. *Applied Sciences*, 11(19). https://doi.org/10.3390/app11198817

Augenstein, Isabelle, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2024. Factuality challenges in the era of large language models and opportunities for fact-checking. *Nature Machine Intelligence*, 6(8):852–863. https://doi.org/10.1038/s42256-024-00881-z

Bai, Jinze, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Bond Jr., Charles F. and Bella M. DePaulo. 2006. Accuracy of deception judgments. *Personality and Social Psychology Review*, 10(3):214–234. https://doi.org/10.1207/s15327957pspr1003_2, PubMed: 16859438

Bonferroni, Carlo. 1936. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62.

Burgoon, Judee K., J. Pete Blair, Tiantian Qin, and Jay F. Nunamaker. 2003. Detecting deception through linguistic analysis. In *Intelligence and Security Informatics: First NSF/NIJ Symposium, ISI 2003*, pages 91–101. https://doi.org/10.1007/3-540-44853-5_7

Burgoon, Judee K. and Tiantian Qin. 2006. The dynamic nature of deceptive verbal communication. *Journal of Language and Social Psychology*, 25(1):76–96. https://doi.org/10.1177/0261927X05284482

Capuozzo, Pasquale, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020a. Automatic detection of cross-language verbal deception. *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

Capuozzo, Pasquale, Ivano Lauriola, Carlo Strapparava, Fabio Aiolli, and Giuseppe Sartori. 2020b. DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1423–1430.

Chuang, Yun-Shiuan, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346. https://doi.org/10.18653/v1/2024.findings-naacl.211

Crudu, Mălina and Mihaiela Lupea. 2024. Detecting textual deception in courtroom testimonies for Romanian language using neural network methods. In *2024 IEEE 20th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 1–8. https://doi.org/10.1109/ICCP63557.2024.10793021

Dando, Coral J., Alexandra L. Sandham, Charlotte Sibbons, and Paul J. Taylor. 2024. Arabic within culture forensic interviews: Arabic native speaking lay-observer truth and lie accuracy, confidence, and verbal cue selection. *PLOS ONE*, 19(9):1–21. https://doi.org/10.1371/journal.pone.0310384, PubMed: 39312557

Dando, Coral J., Paul J. Taylor, and Alexandra L. Sandham. 2023. Cross cultural verbal cues to deception: Truth and lies in first and second language forensic interview contexts. *Frontiers in Psychology*, 14:1152904. https://doi.org/10.3389/fpsyg.2023.1152904, PubMed: 37325746

DeAndrea, David C., Stephanie Tom Tong, Yuhua Jake Liang, Timothy R. Levine, and Joseph B. Walther. 2012. When do people misrepresent themselves to others? The effects of social desirability, ground truth, and accountability on deceptive self-presentations. *Journal of*

Communication, 62(3):400–417. https://doi.org/10.1111/j.1460-2466.2012.01646.x

DePaulo, Bella M., James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. 2003. Cues to deception. *Psychological Bulletin*, 129(1):74–118. https://doi.org/10.1037/0033-2909.129.1.74, PubMed: 12555795

de Ruiter, Bob and George Kachergis. 2018. The Mafiascum dataset: A large text corpus for deception detection. *ArXiv*, abs/1811.07851.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Elliot, Andrew J. and Patricia G. Devine. 1994. On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3):382.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press. https://doi.org/10.1515/9781503620766

Fornaciari, Tommaso, Federico Bianchi, Massimo Poesio, and Dirk Hovy. 2021. BERTective: Language models and contextual information for deception detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2699–2708. https://doi.org/10.18653/v1/2021.eacl-main.232

Fornaciari, Tommaso and Massimo Poesio. 2012. DeCour: A corpus of DEceptive statements in Italian COURts. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1585–1590. https://doi.org/10.63317/5k56ozq5tzn5

Fornaciari, Tommaso and Massimo Poesio. 2013. Automatic deception detection in Italian court cases. *Artificial Intelligence and Law*, 21(3):303–340. https://doi.org/10.1007/s10506-013-9140-4

Fornaciari, Tommaso and Massimo Poesio. 2014. Identifying fake Amazon reviews as learning from crowds. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 279–287. https://doi.org/10.3115/v1/E14-1030

Glass, G. V. and K. D. Hopkins. 1996. *Statistical Methods in Education and Psychology*. Allyn and Bacon.

Goffman, Erving. 1959. *The Presentation of Self in Everyday Life*. Anchor Books.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 herd of models.

Harmon-Jones, Eddie. 2000. A cognitive dissonance theory perspective on the role of emotion in the maintenance and change of beliefs and attitudes. In N. H. Frijda, A. S. R. Manstead, and S. Bem, editors, *Emotions and Belief: How Feelings Influence Thoughts*. Cambridge University Press, pages 185–211. https://doi.org/10.1017/CBO9780511659904.008

Hauch, Valerie, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L. Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review*, 19(4):307–342. https://doi.org/10.1177/1088868314556539, PubMed: 25387767

Hazra, Sanchaita and Bodhisattwa Prasad Majumder. 2024. To tell the truth: Language of deception and language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8506–8520. https://doi.org/10.18653/v1/2024.naacl-long.470

Hernández-Castañeda, Ángel, Hiram Calvo, Alexander Gelbukh, and Jorge J. García Flores. 2016. Cross-domain deception detection using support vector networks. *Soft Computing*, 21(3):585–595. https://doi.org/10.1007/s00500-016-2409-2

Jaume Masip, Siegfried L. Sporer, Eugenio Garrido and Carmen Herrero. 2005. The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime & Law*, 11(1):99–122. https://doi.org/10.1080/1068316041000172635

Jiang, Albert Q., Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, et al. 2024. Mixtral of experts.

Johnson, Marcia K. and Carol L. Raye. 1981. Reality monitoring. *Psychological Review*, 88(1):67. https://doi.org/10.1037/0033-295X.88.1.67

Kang, Seung Man and Hyoungkeun Lee. 2014. Detecting deception by analyzing written statements in Korean. *Linguistic Evidence in Security, Law and Intelligence*, 2(2). https://doi.org/10.5195/lesli.2014.13

Kincaid, J. Peter, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, Fog count and Flesch Reading Ease formula) for Navy enlisted personnel. Technical Report 8–75, University of Central Florida, Institute for Simulation and Training. https://doi.org/10.21236/ADA006655

Kleinberg, Bennett, Isabelle van der Vegt, Arnoud Arntz, and Bruno Verschuere. 2019. Detecting deceptive communication through linguistic concreteness. *PsyArxiv Preprints*. https://doi.org/10.31234/osf.io/p3qjh

Kleinberg, Bennett and Bruno Verschuere. 2021. How humans impair automated deception detection performance. *Acta Psychologica*, 213:103250. https://doi.org/10.1016/j.actpsy.2020.103250, PubMed: 33450692

Leal, Sharon, Aldert Vrij, Zarah Vernham, Gary Dalton, Louise Jupe, Adam Harvey, and Galit Nahari. 2018. Cross-cultural verbal deception. *Legal and Criminological Psychology*, 23(2):192–213. https://doi.org/10.1111/lcrp.12131

Levine, Timothy R. 2014. Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, 33(4):378–392. https://doi.org/10.1177/0261927X14535916

Li, Deqing and Eugene Santos. 2012. Argument formation in the reasoning process: Toward a generic model of deception detection. In *Proceedings of the Workshop on Computational Approaches to Deception Detection*, pages 63–71.

Li, Haipeng and Jonathan Dunn. 2022. Corpus similarity measures remain robust across diverse languages. *Lingua*, 275:103377. https://doi.org/10.1016/j.lingua.2022.103377

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252. https://doi.org/10.18653/v1/2022.acl-long.229

Loconte, Riccardo, Roberto Russo, Pasquale Capuozzo, Pietro Pietrini, and Giuseppe Sartori. 2023. Verbal lie detection using large language models. *Scientific Reports*, 13:22849. https://doi.org/10.1038/s41598-023-50214-0, PubMed: 38129677

Mahon, James Edwin. 2016. The definition of lying and deception. *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), edited by Edward N. Zalta.

McCornack, Steven, Leslie Baxter, and Dawn Braithwaite. 2008. Information manipulation theory: Explaining how deception occurs. In *Engaging Theories in Interpersonal Communication: Multiple Perspectives*. SAGE Publications, Inc., pages 215–226. https://doi.org/10.4135/9781483329529.n16

McCornack, Steven A. 1992. Information manipulation theory. *Communication Monographs*, 59(1):1–16. https://doi.org/10.1080/03637759209376245

Mihalcea, Rada and Carlo Strapparava. 2009. The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, ACLShort '09, pages 309–312. https://doi.org/10.3115/1667583.1667679

Nahari, Galit and Aldert Vrij. 2019. The verifiability approach: Advances, challenges, and future prospects. *The Routledge International Handbook of Legal and Investigative Psychology*, pages 212–223. https://doi.org/10.4324/9780429326530-15

Nahari, Galit, Aldert Vrij, and Ronald P. Fisher. 2014. Exploiting liars' verbal strategies by examining the verifiability of details. *Legal and Criminological Psychology*, 19(2):227–239. https://doi.org/10.1111/j.2044-8333.2012.02069.x

Newman, Matthew L., James W. Pennebaker, Diane S. Berry, and Jane M. Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, 29(5):665–675. https://doi.org/10.1177/0146167203029005010, PubMed: 15272998

Ott, Myle, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 497–501.

Ott, Myle, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319.

Papantoniou, Katerina, Panagiotis Papadakos, Theodore Patkos, George Flouris, Ion Androutsopoulos, and Dimitris Plexousakis. 2022. Deception detection in text and its relation to the cultural dimension of individualism/collectivism. *Natural Language Engineering*, 28(5):545–606. https://doi.org/10.1017/S1351324921000152

Pennebaker, James W., Ryan Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The Development and Psychometric Properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z

Pennebaker, J. W. and C. K. Chung. 2001. Expressive writing: Connections to physical and mental health. *Advances in Experimental Social Psychology*, 34:41–61.

Pérez-Rosas, Verónica, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 59–66. https://doi.org/10.1145/2818346.2820758

Pérez-Rosas, Verónica and Rada Mihalcea. 2014. Cross-cultural deception detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 440–445. https://doi.org/10.3115/v1/P14-2072

Pérez-Rosas, Verónica and Rada Mihalcea. 2015. Experiments in open domain deception detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1120–1125. https://doi.org/10.18653/v1/D15-1133

Peskov, Denis, Benny Cheng, Ahmed Elgohary, Joe Barrow, Cristian Danescu-Niculescu-Mizil, and Jordan Boyd-Graber. 2020. It takes two to lie: One to lie, and one to listen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3811–3854. https://doi.org/10.18653/v1/2020.acl-main.353

Pisarevskaya, Dina, Tatiana Litvinova, and Olga Litvinova. 2017. Deception detection for the Russian language: Lexical and syntactic parameters. In *Proceedings of the 1st Workshop on Natural Language Processing and Information Retrieval associated with RANLP 2017*, pages 1–10. https://doi.org/10.26615/978-954-452-038-0_001

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Salvetti, Franco, John B. Lowe, and James H. Martin. 2016. A tangled web: The faint signals of deception in text - Boulder Lies and Truth Corpus (BLT-C). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3510–3517 https://doi.org/10.63317/299qy7u4dwux

Sap, Maarten, Eric Horvitz, Yejin Choi, Noah A. Smith, and James Pennebaker. 2020. Recollection versus imagination: Exploring human memory and cognition via neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1970–1978. https://doi.org/10.18653/v1/2020.acl-main.178

Skalicky, Stephen Cameron, Nicholas D. Duran, and Scott Andrew Crossley. 2020. Please, please, just tell me: The linguistic features of humorous deception. *Dialogue Discourse*, 11:128–149. https://doi.org/10.5210/dad.2020.205

Soldner, Felix, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777. https://doi.org/10.18653/v1/N19-1175

Sporer, Siegfried L. 2004. Reality monitoring and detection of deception. In P.-A. Granhag and L. Strömwall, editors, *The Detection of Deception in Forensic Contexts*. Cambridge University Press, pages 64–101. https://doi.org/10.1017/CBO9780511490071.004

Sporer, Siegfried Ludwig. 1997. The less travelled road to truth: Verbal cues in deception detection in accounts of fabricated and self-experienced events. *Applied Cognitive Psychology*, 11(5):373–397.

https://doi.org/10.1002/(SICI)1099-0720(199710)11:5<373::AID-ACP461>3.0.CO;2-0

Tache, Irina, Lara Warmelink, Paul Taylor, and Lorraine Hope. 2023. Cultural differences in the efficacy of unexpected questions, sketching, and timeline methods in eliciting cues to deception. *Frontiers in Psychology*, 14:1175333. https://doi.org/10.3389/fpsyg.2023.1175333, PubMed: 37720643

Taylor, Paul J., Samuel Larner, Stacey M. Conchie, and Sophie Van der Zee. 2014. Cross-cultural deception detection. *Detecting Deception: Current Challenges and Cognitive Approaches*. pages 175–201. https://doi.org/10.1002/9781118510001.ch8

Temnikova, Irina, Silvia Gargova, Ruslana Margova, Veneta Kireva, Ivo Dzhumerov, Tsvetelina Stefanova, and Hristiana Krasteva. 2023. New Bulgarian resources for studying deception and detecting disinformation. In *10th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*.

Ten Brinke, Leanne and Stephen Porter. 2012. Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior*, 36(6):469–477. https://doi.org/10.1037/h0093929, PubMed: 23205594

Tormala, Zachary L. and Richard E. Petty. 2002. What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology*, 83(6):1298–1313. https://doi.org/10.1037/0022-3514.83.6.1298, PubMed: 12500812

Tormala, Zakary L. and Richard E. Petty. 2004. Source credibility and attitude certainty: A metacognitive analysis of resistance to persuasion. *Journal of Consumer Psychology*, 14(4):427–442. https://doi.org/10.1207/s15327663jcp1404_11

Turkle, Sherry. 1995. *Life on the Screen: Identity in the Age of the Internet*. Simon & Schuster Trade.

Turner, Ronny E., Charles Edgley, and Glen Olmstead. 1975. Information control in conversations: Honesty is not always the best policy. *The Kansas Journal of Sociology*, 11(1):69–89. https://doi.org/10.17161/STR.1808.6098

Velutharambath, Aswathy and Roman Klinger. 2023. UNIDECOR: A unified deception corpus for cross-corpus deception detection. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 39–51. https://doi.org/10.18653/v1/2023.wassa-1.5

Velutharambath, Aswathy, Amelie Wührl, and Roman Klinger. 2024a. Can factual statements be deceptive? The DeFaBel corpus of belief-based deception. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2708–2723. https://doi.org/10.63317/4yhfhuvm5dog

Velutharambath, Aswathy, Amelie Wührl, and Roman Klinger. 2024b. How entangled is factuality and deception in German? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9538–9554. https://doi.org/10.18653/v1/2024.findings-emnlp.557

Verhoeven, Ben and Walter Daelemans. 2014. CLiPS Stylometry Investigation (CSI) corpus: A Dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3081–3085. https://doi.org/10.63317/5ihmqb5uqzzs

Vrij, Aldert. 2008. *Detecting Lies and Deceit: Pitfalls and Opportunities*, 2nd edition. Wiley.

Vrij, Aldert, Ronald Fisher, Samantha Mann, and Sharon Leal. 2006. Detecting deception by manipulating cognitive load. *Trends in Cognitive Sciences*, 10(4):141–142. https://doi.org/10.1016/j.tics.2006.02.003, PubMed: 16516533

Vrij, Aldert, Ronald P. Fisher, Samantha Mann, and Sharon Leal. 2010. Lie detection: Pitfalls and opportunities. In G. Daniel Lassiter and Christian A. Meissner, editors, *Police Interrogations and False Confessions: Current Research, Practice, and Policy Recommendations*. American Psychological Association, pages 97–110. https://doi.org/10.1037/12085-006

Walther, Joseph B. 1996. Computer-mediated communication: Impersonal,

interpersonal, and hyperpersonal interaction. *Communication Research*, 23(1):3–43. https://doi.org/10.1177/009365096023001001

Yao, Wenlin, Zeyu Dai, Ruihong Huang, and James Caverlee. 2017. Online deception detection refueled by real world data collection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 793–802. https://doi.org/10.26615/978-954-452-049-6_102

Zuckerman, Miron, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology*, 14:1–59. https://doi.org/10.1016/S0065-2601(08)60369-X