Improving Language Models for Emotion Analysis: Insights from Cognitive Science

Anonymous ACL submission

Abstract

We propose leveraging cognitive science research on emotions and communication to improve language models for emotion analysis. First, we present the main emotion theories in psychology and cognitive science. Then, we introduce the main methods of emotion annotation in natural language processing and their connections to psychological theories. We also present the two main types of analyses of emotional communication in cognitive pragmatics. Finally, based on the cognitive science research presented, we propose directions for improving language models for emotion analysis. We suggest that these research efforts pave the way for constructing new annotation schemes, methods, and a possible benchmark for emotional understanding, considering different facets of human emotion and communication.

1 Introduction

Emotion analysis in natural language processing aims to develop computational models capable of discerning human emotions in text. Recently, language models have been widely used to solve various tasks in natural language processing, including emotion analysis (Devlin et al., 2019; Brown et al., 2020). This field of research faces several limitations. First, different ways of conceptualizing emotions lead to different annotation schemes and datasets (Klinger, 2023). As a result, the generalization ability of models is limited, and it is often impossible to compare studies. To address these limitations, it has been proposed to unify some annotation schemes based on the semantic proximity of emotion categories (Bostan and Klinger, 2018), to automatically find emotion categories from data (De Bruyne et al., 2020), or to obtain emotion embeddings independent of annotation schemes (Buechel et al., 2021). Inspired by psychology and cognitive science research, we believe building an annotation scheme unifying different perspectives on the emotional phenomenon would be possible and desirable.

In addition, existing benchmarks evaluate certain aspects of emotional understanding but do not consider its full complexity (Campagnano et al., 2022; Zhang et al., 2023a; Paech, 2024). For example, Paech (2024) proposes to evaluate the emotional understanding of language models by predicting the intensity of emotions in conflict scenes. This type of evaluation is too limited: benchmarks should reflect as much as possible the richness of emotional understanding in humans, a richness documented in different branches of affective sciences (Green, 2007; Wharton, 2016; Scarantino, 2017; Barrett et al., 2019; Bonard and Deonna, 2023).

Another related research area focuses on the theory of mind of language models, *i.e.*, their ability to correctly attribute mental states to others. In our view, this literature is promising in that it links recent developments in language models to theories and empirical methods in cognitive science (for a review, see Bonard (2024, section 5)). Notably, several tasks and benchmarks have been developed to measure the ability of language models to succeed at different versions of the False Belief Task (Trott et al., 2022; Aru et al., 2023; Gandhi et al., 2023; Holterman and van Deemter, 2023; Kosinski, 2023; Mitchell and Krakauer, 2023; Shapira et al., 2023; Stojnić et al., 2023; Ullman, 2023). However, theory of mind and, more generally, social reasoning abilities go beyond the ability to succeed at the False Belief Task (Apperly and Butterfill, 2009; Langley et al., 2022; Ma et al., 2023). The ability to correctly interpret expressed emotions cannot be reduced to it. The degree to which language models possess this emotional competence is worth studying in its own right.

Generally speaking, research on language models for emotion analysis would benefit from cognitive science research on emotion and communication. In particular, we believe this approach can lead to better ways of annotating emotions expressed in text. Additionally, it can improve the evaluation of the emotional understanding of language models by developing new benchmarks. In what follows, we present an overview of psychological theories of emotion (section 2) and ways of annotating emotions in natural language processing (section 3). Then, inspired by specific psychological and linguistic theories (section 4), we propose research directions to address some of the current limitations of emotion analysis (section 5).

2 Emotion Theories in Cognitive Science

This section will present the three main emotion theories in psychology to provide a background for connecting emotion analysis in natural language processing with cognitive science.

Basic emotion theory. Basic emotion theory is certainly the most influential today. Inspired by Darwin's research on emotions (Darwin, 1872), it postulates a certain number of discrete, basic emotions that are universal and innate among humans due to their evolutionary origins. Emotions are understood as psycho-physiological "programs" that were naturally selected to help overcome recurrent evolutionary challenges (Cosmides and Tooby, 2000). A prominent version is that of Paul Ekman (Ekman, 1999), who sought to show, as Darwin envisaged, that some emotions are expressed with the same facial expressions across cultures – Ekman used Darwin's (Darwin, 1872) list of six "core" expressions of emotions: anger, fear, surprise, disgust, happiness, and sadness. He notably conducted studies with individuals having no exposure to Western culture, indicating that they could accurately identify facial expressions for these six emotions (Ekman and Friesen, 1971). There have also been attempts to support basic emotion theory by identifying physiological and neurological signatures of basic emotions (Moors, 2022, 129—131). It should be noted that Ekman left it open how many basic emotions there are. Besides the six emotions listed, candidates include amusement, contempt, embarrassment, guilt, pride, and shame (Ekman, 1999). Other versions of basic emotion theory have different lists (Tomkins, 1962; Izard, 1992; Panksepp, 1998; Plutchik, 2001).

Psychological constructivism. Psychological constructivism is the most influential alternative to basic emotion theory today. It rejects that there are discrete, basic emotions universally shared by humans and posits instead that emotion kinds such as anger, fear, and joy are constructed through the interplay of biological, psychological, and sociocultural factors. Early proponents include Schachter and Singer (1962), but its main representatives are James Russell and Lisa Feldman Barrett (Russell and Barrett, 1999). Psychological constructivists focus on the feeling component of emotions that they interpret as a continuum with no categorical barriers. Feelings are typically represented in a two-dimensional space with a valence axe (pleasant–unpleasant feelings) and an arousal axe (feelings of activation–deactivation). The impression that there are discrete emotions is seen as a social construct: different forms of enculturation yield different ways to conceptualize or label our bodily feelings into discrete emotional kinds.

Appraisal theory. The third major psychological theory of emotion is appraisal theory, whose empirical version was pioneered by Magda Arnold (Arnold, 1960). It was developed to explain the absence of a bijective, one–to–one correspondence between kinds of emotions and emotional stimuli, *i.e.*, the fact that the same kind of stimuli triggers different emotions and that different kinds of stimuli trigger the same kind of emotion. To explain this fact, appraisals are postulated as mediators between stimuli and emotional reactions. Appraisals are cognitive evaluations (unconscious, fast, and error-prone) of the relevance of stimuli given one's concerns and how one should react. Appraisal theory hypothesizes that, for instance, Sam is fearful of the mouse in the kitchen because he appraises it as an imminent threat to his safety, while Maria, on the other, is angry that there is a mouse in the kitchen because she appraises it as an intruder to be kicked out. Thus, each emotion kind can be analyzed by the associated appraisal. For instance, Lazarus (1991) proposes *imminent danger* for fear, *demeaning offense* for anger, *irrevocable loss* for sadness, and *progress towards a goal* for happiness.

In the 1980s, appraisal theorists started to analyze appraisals as regions in a multi-dimensional space (Moors et al., 2013). Appraisal dimensions typically include (a) the goal-conduciveness of the stimulus, (b) the coping potential of the individual in the situation, (c) the urgency of the needed response, (d) the cause of the eliciting event (me, others, intentional or not), and (e) the compatibility with one's normative standards. For instance, fear is triggered by an appraisal of a stimulus as (a) highly inconducive, (b) hard to cope with, and (c) requiring an urgent response.

An integrated framework for emotion theories. Though the three theories reviewed are usually considered rivals, some have argued for their integration (Scherer and Moors, 2019; Bonard, 2021b; Scherer, 2022). Arguably, the three theories differ mainly in their focus. Basic emotion theory focuses on the universal traits inherited from evolution, particularly their physiological and bodily expressions. Psychological constructivism focuses on the feeling dimensions and how individuals categorize them. Appraisal theory focuses on emotional elicitation and action tendencies. We believe that a framework integrating the various elements studied by these theories is possible and desirable. What we call "the integrated framework for emotion theories" proposes to do so by postulating that paradigmatic emotional episodes are made of synchronized and causally interconnected changes in four components: appraisal process, action tendencies, bodily changes (motor expressions and physiological responses), and subjective feelings. For a discussion of this integrated framework, see Scherer (2022).

3 Emotion Analysis in Text

3.1 How is emotion annotated in text?

Emotion is a category. Textual emotion analysis relies on basic emotion theories to define different emotion categories to associate with textual units (a textual span, a sentence, or a document). For instance, the sentence "I love philosophy." could automatically be associated with the discrete emotion *happiness*. Several annotation schemes focus on subsets of categories while others encompass a broader set, reaching over 28 different categories (Demszky et al., 2020; Bostan and Klinger, 2018).

Emotion is a continuous value with affective meaning. Instead of representing emotion as a category, some annotation schemes consider emotion as a point in a multidimensional space, associating continuous values with textual units (Buechel and Hahn, 2017). These dimensions carry an affective meaning. Two dimensions dominate the literature and stem from psychological constructivism, which considers, as we have seen, that an emotion can be characterized by its degree of *pleasantness* and its degree of *arousal*. Thus, the sentence "His voice soothes me." could be automatically associated with two continuous values: a degree of *pleasantness* of 4 out of 5 and a degree of *arousal* of 1 out of 5.

Emotion is a continuous value with cognitive meaning. These dimensions can also carry a cognitive meaning. Recently, a new line of research proposes incorporating appraisal theories into emotion analysis models (Hofmann et al., 2020; Troiano et al., 2022; Zhan et al., 2023). From this perspective, emotions are caused by events evaluated according to several cognitive dimensions. For example, the sentence "I received a surprise gift." could be automatically associated with several continuous values: the event is *sudden* (4 out of 5), *contrary to social norms* (0 out of 5), and the person has *control* over the event (0 out of 5).

Emotion consists of semantic roles. An emotion cannot be reduced to a category or continuous values with affective or cognitive meaning. To better understand an emotional event, several approaches associate spans of text with semantic roles, such as *cause*, *target*, *experiencer*, and *cue* of the emotion (Lee et al., 2010; Kim and Klinger, 2018; Bostan et al., 2020; Oberländer et al., 2020; Campagnano et al., 2022; Wegge et al., 2023). Thus, instead of considering emotion as caused by an event, semantic role labeling of emotions considers that emotion *is* an event (Klinger, 2023) that must be reconstructed by answering the question: "Who (*experiencer*) feels what (*cue*) towards whom (*target*) and why (*cause*)?". In this example, each text span can be associated with a semantic role: "Louise (*experiencer*) was angry (*cue*) at Paul (*target*) because he did not warn her (*cause*)."

Emotion is a refined feeling. Sentiment analysis, a fundamental task in natural language processing, is sometimes considered a simplified version of emotion analysis. In its most basic form, sentiment analysis associates textual units with a category indicating a polarity (*positive* or *negative*) (Poria et al., 2020). A finer-grained task identifies aspects of a product or topic and determines the sentiment expressed about each of these aspects (Zhang et al., 2022). For example, in the sentence "The battery life of this phone is amazing, but its camera quality is disappointing.", the sentiment is *positive* for the aspect "battery life" and is *negative* for the aspect "camera quality."

3.2 Limitations

No unified annotation scheme. Divergences in the psychological definition of emotion lead to divergences in how emotion is annotated in the text. Psychological theories of emotions represent different perspectives on the emotional phenomenon. However, these perspectives are not as contradictory as they seem and may even tend towards unification (section 2). We believe this is also the case for annotation schemes in emotion analysis. In section 5, we provide directions for constructing a unified annotation scheme inspired by recent debates in psychology (Scherer, 2022).

Emotion verbalization is overlooked. Emotion analysis rarely considers the process of emotion verbalization. As a result, it is difficult to obtain annotation guides that clearly define the linguistic markers to annotate in text. We want to highlight the linguistic theory of Raphael Micheli, which categorizes a broad panel of linguistic markers into three emotion expression modes (Micheli, 2014): *labeled*, *displayed*, and *suggested* emotion. Emotion can be expressed explicitly with an emotional label ("I am *happy* today"), be displayed with linguistic characteristics of an utterance such as interjections and punctuations ("*Ah!* That's great *!*"), or be suggested with the description of a situation that, in a given sociocultural context, leads to an emotion ("*She gave me a gift*"). Most annotation schemes have implicitly focused on the *labeled* emotion, overlooking the other two expression modes. Recently, annotation schemes based on appraisal theories implicitly concern themselves with the *suggested* emotion. Micheli's theory thus analyzes the different types of verbal signs humans use to infer expressed emotions. In a complementary manner, theories of cognitive pragmatics are interested in the psychological mechanisms used to infer what is communicated, especially the emotions expressed by these different types of signs. In the next section, we will hypothesize that the sign categories distinguished by Micheli correspond to different sources of inferences postulated by cognitive pragmatics.

4 Cognitive Pragmatics and Emotional Communication

Two analyses of communication. Cognitive pragmatics is the branch of cognitive science concerned with how agents use and interpret signs in communication. In this and related branches, it is common to distinguish between two broad ways to analyze communication: the "dictionary analysis" (a.k.a. the "code", "semiotic", or "semantic" model) and the "detective analysis" (a.k.a. the "Gricean", "inferential", or "pragmatic" model) (Sperber and Wilson, 1995; Schlenker, 2016; Heintz and Scott-Phillips, 2023).

Dictionary analysis. The dictionary analysis depicts communication as a sender who intentionally or unintentionally encodes information into a signal that the receiver decodes. Vitally, prior to the communicative exchange, the sender and the receiver must share the same code. A code here is understood as a pre-established pairing between kinds of stimuli (symbolized by "<. . .>") and sets of information (symbolized by "[. . .]"). For instance, the Morse code consists of a pairing between <combinations of short and long signals> and [letters] that senders and receivers must share to communicate with it. Codes can be conventional, as the Morse code is and as is the formal semantics of a language: a code made of syntactical and lexical rules that pairs <strings of words> with [sentential meanings] (Heim and Kratzer, 1998). Codes can also be non-conventional or "natural" (Wharton, 2003; Bonard, 2023a). For instance, bees are thought to use a code pairing their <dances> with the [location of nectar]. As mentioned in section 2, humans are thought to use a code pairing types of <facial expressions> with types of [emotions expressed].

The main limitation of the dictionary analysis is that codes sometimes *underdetermine meaning*: The pre-established pairings between <types of stimuli> and [sets of information] are sometimes insufficient to account for the information communicated. Paradigmatically, in *conversational implicatures* (Grice, 1975), the utterer implicitly communicates information beyond what is linguistically encoded, beyond what is determined by syntactical and lexical rules. For instance (Wilson and Sperber, 2006), if Peter asks, "Did John pay back the money he owed you?" and Mary answers, "He forgot to go to the bank.", Peter will readily understand that Mary means "no" although the relevant code – the rules pairing <English grammar and lexicon> with [sentential meaning] – is by itself insufficient to account for this since the code only tells you that John forgot to go to the bank.

Codes underdetermine the meaning of verbal expressions of emotions as well. To illustrate, let us go back to Micheli's typology: *labeled*, *displayed*, and *suggested* emotions (Micheli, 2013). As far as *labeled* emotions are concerned, the dictionary analysis does quite well thanks to the pairing between <emotion words> (*e.g.*, happy, amazing, sadly) with the [emotion kinds] they refer to. However, even *labeled* emotions sometimes do not encode all that is communicated. For instance, "I am happy now" is explicit about the kind of emotion expressed but does not encode what the emotion is about. Nevertheless, we often correctly infer such information in the relevant context. The dictionary analysis fairs even less well with *displayed* emotions because these are often ambiguous. For instance, interjections such as "Wow!", "Damn!", "Fuck!", "Shit!", "Ah!" and "Oh!" though they readily display that the utterer undergoes an emotion, can express various positive and negative emotions. Furthermore, these interjections don't encode what emotions are about. However, receivers usually correctly infer these pieces of information. The dictionary analysis regarding *suggested* emotions is even more limited. Depending on what the person expressing their emotion believes or desires, a phrase that only suggests emotions can communicate pretty much any kind of emotion. Imagine, for instance, that someone says, "The ship has black sails.". In a certain context, this apparently vapid sentence may poignantly convey intense emotion – because, say, it means that the son of the utterer died, as in the story of Aegeus and Theseus. Note that, beyond verbal expression, most, if not all, types of emotional expressions also underdetermine what emotions are expressed. Facial expressions or acoustic cues (*e.g.*, screams, laughter, sighs) also communicate different emotions given different contexts (Aviezer et al., 2008; Teigen, 2008; Vlemincx et al., 2009; Barrett et al., 2011, 2019; Bonard, 2023b). The dictionary analysis is thus also insufficient for these kinds of emotional expressions.

So, how do humans disambiguate emotional expressions in cases where codes underdetermine what is communicated? If we trust contemporary cognitive pragmatics, the answer should be found in the detective analysis of communication.

Detective analysis. What we call the detective analysis is constituted by a family of theories developed by Paul Grice (Grice, 1957, 1989) and his heirs (for reviews, see Bonard (2021a), chapter one and appendix). Note that although our presentation aims to remain balanced, no universally accepted version of this analysis exists.

As mentioned, the detective analysis was developed to account for conversational implicatures, cases where what is communicated goes beyond what is conveyed through conventional meaning, as in Peter and Mary's example above. To do so, the detective analysis conceptualizes linguistic interpretation as a type of abductive reasoning – *i.e.*, as an inference that seeks the simplest and most likely conclusion given the evidence available. The analysis spells out three main sources of evidence:

1. *Codes*, *i.e.*, pre-established pairings between types of stimuli and sets of information, *e.g.*, English syntactical and lexical rules; the codes for verbal and nonverbal emotional expressions. As we saw, expressions using labeled (*e.g.*, "I'm happy") and displayed emotions (*e.g.*, "Damn!") are partially understood through such codes, though they are too ambiguous to account for all that is communicated.

2. *Pragmatic expectations*, *i.e.*, how people are expected to behave in given contexts, particularly the kind of signal they receive. For instance, in conversations, people are expected to say things relevant to the question under discussion (see Grice (1975)'s maxims of conversation). For this reason, although what is literally encoded in Mary's reply is that John forgot to go to the bank, Peter will nevertheless expect this to be relevant to the question he asked. Similarly, we expect someone's emotional expressions to be about something relevant to their concerns (Wharton et al., 2021; Bonard, 2022). For instance, if someone says "Damn!" after receiving a surprisingly nice compliment, we expect the compliment to be particularly relevant to the person and will interpret the interjection accordingly.

3. *Common ground*, *i.e.*, the information presumed to be shared by the participants in the exchange (Stalnaker, 2002). For instance, Mary and Peter both presume that a bank is a place where one can withdraw money. Similarly, we usually presume that receiving a compliment is something that one seeks, especially if it is surprisingly nice – though this is not always part of the common ground, *e.g.*, if the complimenter is the complimentee's arch-

enemy. The common ground also allows us to understand that Aegeus can express deep despair with the sentence « The ship has black sails. ».

Based on these three sources of evidence, the detective analysis further postulates that the interpreter uses *mindreading* abilities (*i.e.*, theory of mind, mentalizing, or social cognition) to infer what is the most likely piece of information that is implicitly communicated – *e.g.*, Peter infers that Mary meant "no" and we infer that the person saying "Damn!" is probably pleased. Finally, the detective analysis specifies that the information so inferred is added to the common ground shared by participants in the exchange so that it may be a new source of evidence in the upcoming exchanges.

Let us note that the detective analysis predicts that the ability to correctly infer what is communicated by emotional expressions heavily depends on one's mind-reading capacities. Corroborating this prediction, children or people on the autistic spectrum may struggle to infer implicit meaning correctly, *e.g.*, conversational implicatures (Foppolo and Mazzaggio, 2024) or in expressions using suggested emotions (Blanc and Quenette, 2017).

5 Research Directions for Emotion Analysis

5.1 Towards a Unified Annotation Scheme

Training models on data annotated with a scheme that reflects the multifaceted nature of emotions is desirable to improve the capacity of language models to understand emotions. Such a scheme would need to integrate different perspectives on the emotional phenomena to allow for better study comparisons. This would also increase the performance and generalization of models.

Attempts at unification. Several recent studies attempt to unify different ways of annotating emotion in text. Campagnano et al. (2022) propose a new annotation scheme that unifies various schemes on emotion semantic roles. To choose a set of shared categories, the different discrete emotions from the schemes were converted to the basic emotions of Plutchik's theory (Plutchik, 2001). Klinger (2023) explores the divergences and commonalities between semantic role labeling of emotions and approaches based on appraisal theories. The study identifies several research directions, such as using appraisal variables to improve the task of detecting emotion causes, or analyzing experiencer-specific appraisals (Wegge et al., 2023). These studies show that combining schemes allows knowledge transfer between tasks, increasing performance and generalization.

In search of a common framework. What we have previously referred to as "the integrated framework for emotion theories" (section 2) aims to reconcile the main emotion theories in psychology (Scherer, 2022). In our view, it represents a strong candidate to provide a common framework for annotation schemes. As mentioned in section 2, this model considers that emotion consists of synchronized changes in different components: the appraisal process, action tendencies, bodily changes (motor expressions and physiological responses), and subjective feelings. Research in emotion analysis must draw from the recent debates in the psychology of emotions to bring existing annotation schemes into dialogue on a solid theoretical basis and, ideally, construct a unified annotation scheme.

Emotion comprises several interacting components. A unified annotation scheme could clarify some gray areas in emotion analysis, such as the lack of clear definitions for emotion semantic roles (*e.g.*, experiencer, cause, and target). It could also better situate existing schemes. For example, annotating discrete emotions and affective dimensions emphasize subjective feeling, whereas annotating cognitive dimensions emphasizes appraisals. Few schemes account for physiological responses, motor expressions, and action tendencies. More generally, few schemes consider all components. Kim and Klinger (2019) analyze the communication of emotions in fiction through descriptions of subjective sensations, postures, facial expressions, and spatial relations between characters. Casel et al. (2021) associate text spans with categories corresponding to Scherer's emotional components. Cortal et al. (2022, 2023) structure emotional narratives according to components similar to Scherer's. Each text span corresponds to observable behaviors, thoughts, physical feelings, or appraisals. To our knowledge, no annotation schemes attempt to capture the interaction between components. Generally, emotion analysis pays little attention to the dynamic nature of emotion and the synchronization of its various components.

Improving the clarity of annotation guides. We note that few studies psychologically justify the choice of different objects to detect in the text. Emotion analysis needs to develop a systematic approach to compare annotation guides with one another, thereby precisely understanding how different annotation schemes capture emotion. Thus, these schemes must draw from psychological theories (section 2) but also from linguistic theories (sections 3.2 and 4) to identify linguistic markers that verbalize emotion. With clear annotation guides, it would be easier for research teams to focus on points of convergence between schemes.

5.2 Better Knowledge Use and Environmental Interaction

In natural language processing, *prompting* refers to supplying a tailored input to a language model, aiming to direct its generation process towards a desired response (Brown et al., 2020). Numerous prompting methods draw inspiration from human cognition to improve the performance of language models (Zhang et al., 2023b). These methods propose generating reasoning steps (Wei et al., 2023; Kojima et al., 2023), reasoning through multiple generated responses (Wang et al., 2023b; Yoran et al., 2023), facilitating communication by rephrasing questions (Deng et al., 2023), and self-improving with its own generated feedback (Madaan et al., 2023; Yuan et al., 2024).

Prompting methods for emotional understanding. Most methods have been explored to improve model performance on tasks requiring formal reasoning (Zhang et al., 2023b). We believe it is possible to adapt these methods or even create new ones to improve model performance on tasks requiring social reasoning, such as emotional understanding. It would be interesting to rely on the ability of language models to act as character simulators (Shanahan et al., 2023; Lu et al., 2024), capable of adopting multiple perspectives to change style (Deshpande et al., 2023), solve tasks requiring expert knowledge (Xu et al., 2023), or simulate discussions to encourage exploration (Wang et al., 2023c; Liang et al., 2023). Zhou et al. (2023) enhance the ability of language models to make relevant inferences for solving theory of mind tasks. They propose a reasoning structure that anticipates future challenges and reasons about potential actions. More globally, a major challenge in natural language processing is finding suitable reasoning structures to effectively use the internal knowledge of models (Kojima et al., 2023; Zhou et al., 2023, 2024). The contribution of the detective analysis (section 4) could prove valuable here: prompts that explicitly ask models to seek evidence from the three sources highlighted by this analysis could lead to better performance and explainability. Finally, the integrated framework for emotion theories (section 3) can serve as inspiration for prompts that aim to exploit all the different facets of emotions rather than focusing on just one of them (*e.g.*, subjective feeling).

Interaction with the environment. Current language models, trained solely on predicting missing words, have essentially mastered linguistic codes, *i.e.*, lexical and syntactic rules (section 4), which Mahowald et al. (2023) call "formal linguistic competence". However, they struggle to perform well on tasks relying on what Mahowald et al. (2023) call "functional linguistic competence", *i.e.* the skills required to use language in real-world situations. These skills centrally involve the mechanisms postulated by the detective analysis – in particular, sharing a common ground and having sensible pragmatic expectations (section 4). To address this limitation, studies augment language models with external modules like a mathematical calculator (Schick et al., 2023), a web browser (Gur et al., 2023), or a virtual environment (Park et al., 2023). Through tool manipulation, language models intertwine reasoning with action and can thus effectively combine internal with external knowledge (Yao et al., 2023). This point is crucial to develop models that exhibit human-like social behaviors. For example, Park et al. (2023) show that observation, planning, and reflection are important components for increasing the credibility of behaviors in a virtual environment. Research on human communication can help highlight relevant abilities to augment language models (*e.g.*, with external modules). This surely applies to emotional communication as well.

5.3 Language Models for Emotion Regulation

Regulating one's emotions and those of others is a fundamental element of emotional intelligence (Mayer et al., 2008; O'Connor et al., 2019). Recently, studies propose assisting psychotherapies with language models (Ziems et al., 2022; Cortal et al., 2022, 2023; Sharma et al., 2023; Chen et al., 2023) to address some public health problems, such as the shortage of mental health professionals, as well as the high cost and social stigma associated with consultations (White and Dorman, 2001). Ziems et al. (2022) perform style transfer to positively reframe negative thoughts according to strategies from positive psychology. To generate positive perspectives that preserve the content of a thought, the study relies on the detective analysis and, more specifically, on Grice's conversational implicatures (section 4). Studies propose automating specific steps of cognitive-behavioral therapies (Beck, 1976) by detecting cognitive distortions (Chen et al., 2023) or performing cognitive reframing (Sharma et al., 2023). We believe the ability of language models to simulate new perspectives on events could be exploited for emotion regulation. As previously seen, it would be possible to automatically provide individuals with new ways of seeing and acting on the world. Such a task would benefit from the knowledge acquired in cognitive pragmatics (section 4).

5.4 Better Benchmarks for Emotional Understanding

Recent benchmarks evaluate language models on specific aspects of emotional understanding (Wang et al., 2023a; Paech, 2024), but they don't consider its full richness (Scherer, 2007; Mayer et al., 2008; O'Connor et al., 2019). For example, Paech (2024) assesses emotional understanding by predicting the intensity of multiple emotions in conflict scenes. Some benchmarks evaluate models on related tasks, such as sentiment analysis (Zhang et al., 2023a) and theory of mind (Zhou et al., 2023; Ma et al., 2023; Kim et al., 2023; Gandhi et al., 2023). However, no benchmark specifically proposes to evaluate the multiple facets of emotions that affective sciences reveal (section 2). Therefore, it is difficult to know whether current models are efficient for emotional understanding.

This limitation is compounded by the fact that it is difficult to clearly determine which properties of emotional understanding are to be evaluated. We believe that evaluating language models should be grounded in research on human emotional communication, especially psycholinguistics. For example, before the age of ten, basic emotions (e.g., joy or sadness) are better remembered than complex emotions (e.g., pride or guilt) (Davidson et al., 2001; Creissen and Blanc, 2017). From six to ten years old, *labeled* emotions are better understood than *suggested* emotions (Blanc, 2010; Creissen and Blanc, 2017). Another example of relevant studies concerns the difficulty that autistic people have in understanding different types of emotional expressions (Foppolo and Mazzaggio, 2024). These studies show that, for humans, different types of emotions and different modes of emotional expression are more or less difficult to interpret. It would be desirable for benchmarks to evaluate language models in ways that reflect the relative difficulty of tasks for humans. Such a project would certainly benefit from research in cognitive pragmatics (section 4), knowing, for example, that people with communication disorders have difficulty understanding conversational implicatures (Foppolo and Mazzaggio, 2024), which indicates that the different sources of evidence distinguished by the detective analysis are associated with different levels of difficulty.

We believe the concept of emotion should be addressed through its relationship with text understanding, i.e., the ability of a reader to construct a mental representation of a situation in a text (Zwaan and Radvansky, 1998). Thus, we would need to go beyond current conceptualizations of emotion in natural language processing (section 3.1) to consider the diversity of linguistic markers used to verbalize emotion (section 3.2) as well as the different types of emotion (basic or complex) from psycholinguistic research (section 2). Inspired by previous studies, Etienne et al. (2022) propose an annotation scheme that considers emotion expression modes and types of emotion. Future benchmarks assessing the ability of language models to analyze emotions should consider such annotation schemes, which, as we have recommended, seek to be solidly based on relevant research in cognitive science.

6 Conclusion

Emotion analysis has several limitations that, we believe, are partially due to a lack of communication with other disciplines and, in particular, cognitive science. We propose exploiting cognitive science research on emotions and communication to address some of these limitations. We suggest that this opens the way for constructing new annotation schemes, methods, and benchmarks for emotional understanding that consider the multiple facets of human emotion and communication.

Limitations

We propose a theoretical perspective on emotion analysis in natural language processing. We believe it would benefit the emotion analysis community to adopt an interdisciplinary approach by drawing from cognitive science theories to address certain existing limitations in the research field. In practice, this is a challenging task. Although we focus on concrete actions that could be undertaken soon (for example, clarifying annotation guidelines), we recognize that our contribution involves speculative research directions. In future research, it would be desirable to complement these speculative aspects with more concrete proposals, notably with empirically testable hypotheses and implementable algorithms.

Ethics Statement

We have not conducted any experimentation or published any data or models in this paper. The present research aims to better understand human emotional communication, not to develop tools for automatically detecting individuals' private subjective states. While we believe our paper does not present direct ethical concerns, the research directions it raises could indirectly harm individuals and societal structures. Although we have highlighted the potential benefits of natural language processing applications (such as emotion regulation tools), it is crucial to ensure that the development and use of such tools do not have any adverse effects in the future.

Acknowledgements

References

Ian A. Apperly and Stephen A. Butterfill. 2009. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953. Publisher: American Psychological Association.

Magda B. Arnold. 1960. *Emotion and Personality*. Columbia University Press, New York.

Jaan Aru, Aqeel Labash, Oriol Corcoll, and Raul Vicente. 2023. Mind the gap: challenges of deep learning approaches to Theory of Mind. *Artificial Intelligence Review*, 56(9):9141–9156.

Hillel Aviezer, Ran R. Hassin, Jennifer Ryan, Cheryl Grady, Josh Susskind, Adam Anderson, Morris Moscovitch, and Shlomo Bentin. 2008. Angry, disgusted, or afraid? Studies on the malleability of emotion perception. *Psychological science*, 19(7):724–732. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*. Publisher: SAGE PublicationsSage CA: Los Angeles, CA.

Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Aaron T. Beck. 1976. *Cognitive Therapy and the Emotional Disorders*. Cognitive Therapy and the Emotional Disorders. International Universities Press, Oxford, England.

Nathalie Blanc. 2010. La compréhension des contes entre 5 et 7 ans: Quelle représentation des informations émotionnelles? [The comprehension of the tales between 5 and 7 year-olds: Which representation of emotional information?]. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, 64(4):256–265.

Nathalie Blanc and Guy Quenette. 2017. La production d'inférences émotionnelles entre 8 et 10 ans: quelle méthodologie pour quels résultats? *Enfance*, 4(4):503–511. Publisher: NecPlus.

Constant Bonard. 2021a. *Meaning and emotion: The extended Gricean model and what emotional signs mean*. Doctoral dissertation, University of Geneva and University of Antwerp.

Constant Bonard. 2021b. Émotions et sensibilité aux valeurs : quatre conceptions philosophiques contemporaines. *Revue de métaphysique et de morale*, 110(2):209–229. Place: Paris cedex 14 Publisher: Presses Universitaires de France.

Constant Bonard. 2022. Beyond ostension: Introducing the expressive principle of relevance. *Journal of Pragmatics*, 187:13–23.

Constant Bonard. 2023a. Natural meaning, probabilistic meaning, and the interpretation of emotional signs. *Synthese*, 201(5):167. Publisher: Springer.

Constant Bonard. 2023b. Underdeterminacy without ostension: A blind spot in the prevailing models of communication. *Mind & Language*.

Constant Bonard. 2024. Can AI and humans genuinely communicate? In Anna Strasser, editor, *Anna's AI Anthology. How to live with smart machines?* Xenemoi, Berlin.

Constant Bonard and Julien Deonna. 2023. Emotion and language in philosophy. In Gesine Lenore Schiewer, Jeanette Altarriba, and Bee Chin Ng, editors, *Language and emotion: An international handbook*, volume 1, pages 54–72. de Gruyter, Berlin.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 1554–1566, Marseille, France. European Language Resources Association.

Laura-Ana-Maria Bostan and Roman Klinger. 2018. An analysis of annotated corpora for emotion classification in text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2104–2119, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel, Luise Modersohn, and Udo Hahn. 2021. Towards label-agnostic emotion embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9231–9249, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cesare Campagnano, Simone Conia, and Roberto Navigli. 2022. SRL4E – Semantic Role Labeling for Emotions: A unified evaluation framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4586–4601, Dublin, Ireland. Association for Computational Linguistics.

Felix Casel, Amelie Heindl, and Roman Klinger. 2021. Emotion recognition under consideration of the emotion component process model. In *Proceedings of the 17th Conference on Natural Language Processing (KONVENS 2021)*, pages 49–61, Düsseldorf, Germany. KONVENS 2021 Organizers.

Zhiyu Chen, Yujie Lu, and William Wang. 2023. Empowering Psychotherapy with Large Language Models: Cognitive Distortion Detection through Diagnosis of Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2022. Natural Language Processing for Cognitive Analysis of Emotions.

Gustave Cortal, Alain Finkel, Patrick Paroubek, and Lina Ye. 2023. Emotion recognition based on psychological components in guided narratives for emotion regulation. In *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 72–81, Dubrovnik, Croatia. Association for Computational Linguistics.

Leda Cosmides and John Tooby. 2000. Evolutionary psychology and the emotions. In Michael Lewis and Jeannette M. Haviland-Jones, editors, *Handbook of emotions*, 2nd edition, pages 91–115. Guilford Press, New York. Publisher: Citeseer.

S. Creissen and N. Blanc. 2017. Quelle représentation des différentes facettes de la dimension émotionnelle d'une histoire entre l'âge de 6 et 10 ans ? Apports d'une étude multimédia. *Psychologie Française*, 62(3):263–277.

Charles Darwin. 1872. *The expression of the emotions in man and animals*. John Murray, London.

Denise Davidson, Zupei Luo, and Matthew J. Burden. 2001. Children's recall of emotional behaviours, emotional labels, and nonemotional behaviours: Does emotion enhance memory? *Cognition and Emotion*, 15(1):1–26.

Luna De Bruyne, Orphee De Clercq, and Veronique Hoste. 2020. An emotional mess! deciding on a framework for building a Dutch emotion-annotated corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1643–1651, Marseille, France. European Language Resources Association.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

Yihe Deng, Weitong Zhang, Zixiang Chen, and Quanquan Gu. 2023. Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves.

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. Toxicity in ChatGPT: Analyzing Persona-assigned Language Models.

Jacob Devlin, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Ekman. 1999. Basic emotions. In Tim Dalgleish and Mike J. Power, editors, *Handbook of cognition and emotion*, pages 45–60. John Wiley & Sons Ltd, Chichester.

Paul Ekman and W V Friesen. 1971. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17 2:124–9.

Aline Etienne, Delphine Battistelli, and Gwénolé Lecorvé. 2022. A (psycho-)linguistically motivated scheme for annotating and exploring emotions in a genre-diverse corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 603–612, Marseille, France. European Language Resources Association.

Francesca Foppolo and Greta Mazzaggio. 2024. Conversational Implicature and Communication Disorders. In Martin J. Ball, Nicole Müller, and Elizabeth Spencer, editors, *The Handbook of Clinical Linguistics, Second Edition*, 1 edition, pages 15–27. Wiley.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. 2023. Understanding Social Reasoning in Language Models with Language Models.

Mitchell Green. 2007. *Self-expression*. Oxford University Press, Oxford.

H. Paul Grice. 1957. Meaning. *The Philosophical Review*, 66(3):377–388.

H. Paul Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill, Leiden.

H. Paul Grice. 1989. *Studies in the way of words*. Harvard University Press, Cambridge (MA).

Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. A Real-World WebAgent with Planning, Long Context Understanding, and Program Synthesis.

Irene Heim and Angelika Kratzer. 1998. *Semantics in generative grammar*. Wiley, Hoboken. Google-Books-ID: jAvR2DB3pPIC.

Christophe Heintz and Thom Scott-Phillips. 2023. Expression unleashed: The evolutionary & cognitive foundations of human communication. *Behavioral and Brain Sciences*, 46:E1. Type: article.

Jan Hofmann, Enrica Troiano, Kai Sassenberg, and Roman Klinger. 2020. Appraisal theories for emotion classification in text. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 125–138, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bart Holterman and Kees van Deemter. 2023. Does ChatGPT have Theory of Mind? ArXiv:2305.14020 [cs].

Carroll E. Izard. 1992. Basic Emotions, Relations Among Emotions, and Emotion-Cognition Relations. *Psychological Review*, 99(3):561–565.

Evgeny Kim and Roman Klinger. 2018. Who Feels What and Why? Annotation of a Literature Corpus with Semantic Roles of Emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Evgeny Kim and Roman Klinger. 2019. An analysis of emotion communication channels in fan-fiction: Towards emotional storytelling. In *Proceedings of the Second Workshop on Storytelling*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A Benchmark for Stress-testing Machine Theory of Mind in Interactions.

Roman Klinger. 2023. Where are We in Event-centric Emotion Analysis? Bridging Emotion Role Labeling and Appraisal-based Approaches. In *Proceedings of the Big Picture Workshop*, pages 1–17, Singapore. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners.

Michal Kosinski. 2023. Theory of Mind Might Have Spontaneously Emerged in Large Language Models. ArXiv:2302.02083 [cs].

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J. Sahakian. 2022. Theory of Mind and Preference Learning at the Interface of Cognitive Science, Neuroscience, and AI: A Review. *Frontiers in Artificial Intelligence*, 5.

Richard S. Lazarus. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American psychologist*, 46(8):819.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large Language Models are Superpositions of All Characters: Attaining Arbitrary Role-play via Self-Alignment.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards A Holistic Landscape of Situated Theory of Mind in Large Language Models.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2023. Dissociating language and thought in large language models.

John D. Mayer, Richard D. Roberts, and Sigal G. Barsade. 2008. Human Abilities: Emotional Intelligence. *Annual Review of Psychology*, 59(1):507–536.

Raphaël Micheli. 2013. Esquisse d'une typologie des différents modes de sémiotisation verbale de l'émotion. *Semen*, (35).

Raphaël Micheli. 2014. *Les émotions dans les discours*. De Boeck Supérieur.

Melanie Mitchell and David C. Krakauer. 2023. The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120. Publisher: Proceedings of the National Academy of Sciences.

Agnes Moors. 2022. *Demystifying emotions: A Typology of theories in psychology and philosophy*, cambridge university press edition. Cambridge.

Agnes Moors, Phoebe C. Ellsworth, Klaus R. Scherer, and Nico H. Frijda. 2013. Appraisal theories of emotion: state of the art and future development. *Emotion Review*, 5(2):119–124. Publisher: Sage Publications Sage UK: London, England.

Laura Oberländer, Kevin Reich, and Roman Klinger. 2020. Experiencers, Stimuli, or Targets: Which Semantic Roles Enable Machine Learning to Infer the Emotions? *arXiv:2011.01599 [cs]*.

Peter J. O'Connor, Andrew Hill, Maria Kaya, and Brett Martin. 2019. The measurement of emotional intelligence: A critical review of the literature and recommendations for researchers and practitioners. *Frontiers in psychology*, 10:1116. Publisher: Frontiers.

Samuel J. Paech. 2024. EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models.

Jaak Panksepp. 1998. *Affective neuroscience: the foundations of human and animal emotions*. Oxford University Press, New York.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior.

Robert Plutchik. 2001. The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, and Rada Mihalcea. 2020. Beneath the Tip of the Iceberg: Current Challenges and New Directions in Sentiment Analysis Research.

James A. Russell and Lisa Barrett. 1999. Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of personality and social psychology*, 76:805–19.

Andrea Scarantino. 2017. How to do things with emotional expressions: The theory of affective pragmatics. *Psychological Inquiry*, 28(2-3):165–185. Publisher: Taylor & Francis.

Stanley Schachter and Jerome Singer. 1962. Cognitive, social, and physiological determinants of emotional state. *Psychological review*, 69(5):379. Publisher: American Psychological Association.

Klaus R. Scherer. 2007. Componential emotion theory can inform models of emotional competence. Publisher: Oxford University Press.

Klaus R. Scherer. 2022. Theory convergence in emotion science is timely and realistic. *Cognition and Emotion*, 36(2):154–170.

Klaus R. Scherer and Agnes Moors. 2019. The emotion process: event appraisal and component differentiation. *Annual Review of Psychology*, 70:719–745. Publisher: Annual Reviews.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools.

Philippe Schlenker. 2016. The semantics-pragmatics interface. In Maria Aloni and Paul Dekker, editors, *The Cambridge Handbook of Formal Semantics*, pages 664–727. Cambridge University Press, Cambridge.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023. Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models. ArXiv:2305.14763 [cs].

Ashish Sharma, Kevin Rushton, Inna Wanyin Lin, David Wadden, Khendra G. Lucas, Adam S. Miner, Theresa Nguyen, and Tim Althoff. 2023. Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and cognition*, 2nd edition edition. Blackwell, Oxford and Cambridge (MA).

Robert Stalnaker. 2002. Common ground. *Linguistics and philosophy*, 25(5/6):701–721.

Gala Stojnić, Kanishk Gandhi, Shannon Yasuda, Brenden M. Lake, and Moira R. Dillon. 2023. Commonsense psychology in human infants and machines. *Cognition*, 235:105406.

Karl Halvor Teigen. 2008. Is a sigh "just a sigh"? Sighs as emotional signals and responses to a difficult task. *Scandinavian journal of Psychology*, 49(1):49–57. Publisher: Wiley Online Library.

Silvan Tomkins. 1962. *Affect imagery consciousness*, volume Volume I: The positive affects. Springer, New York.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2022. Dimensional Modeling of Emotions in Text with Appraisal Theories: Corpus Creation, Annotation Reliability, and Prediction. *Computational Linguistics*, pages 1–71.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2022. Do Large Language Models know what humans know? *arXiv preprint arXiv:2209.01515*.

Tomer Ullman. 2023. Large Language Models Fail on Trivial Alterations to Theory-of-Mind Tasks. ArXiv:2302.08399 [cs].

Elke Vlemincx, Ilse Van Diest, Steven De Peuter, Johan Bresseleers, Katleen Bogaerts, Stien Fannes, Wan Li, and Omer Van Den Bergh. 2009. Why do you sigh? Sigh rate during induced stress and relief. *Psychophysiology*, 46(5):1005–1013. Publisher: Wiley Online Library.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023a. Emotional intelligence of Large Language Models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-Consistency Improves Chain of Thought Reasoning in Language Models.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023c. Unleashing Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration.

Maximilian Wegge, Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Experiencer-Specific Emotion and Appraisal Prediction.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Tim Wharton. 2003. Natural pragmatics and natural codes. *Mind & language*, 18(5):447–477. Publisher: Wiley Online Library.

Tim Wharton. 2016. That bloody so-and-so has retired: Expressives revisited. *Lingua*, 175:20–35. Publisher: Elsevier.

Tim Wharton, Constant Bonard, Daniel Dukes, David Sander, and Steve Oswald. 2021. Relevance and emotion. *Journal of Pragmatics*, 181:259–269.

M. White and S. M. Dorman. 2001. Receiving social support online: Implications for health education. *Health Education Research*, 16(6):693–707.

Deirdre Wilson and Dan Sperber. 2006. Relevance theory. In Laurence Horn, editor, *The Handbook of pragmatics*. Blackwell, Oxford.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models.

Ori Yoran, Tomer Wolfson, Ben Bogin, Uri Katz, Daniel Deutch, and Jonathan Berant. 2023. Answering Questions by Meta-Reasoning over Multiple Chains of Thought. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5942–5966, Singapore. Association for Computational Linguistics.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-Rewarding Language Models.

Hongli Zhan, Desmond Ong, and Junyi Jessy Li. 2023. Evaluating Subjective Cognitive Appraisals of Emotions from Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14418–14446, Singapore. Association for Computational Linguistics.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023a. Sentiment Analysis in the Era of Large Language Models: A Reality Check.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A Survey on Aspect-Based Sentiment Analysis: Tasks, Methods, and Challenges. *arXiv:2203.01054 [cs]*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, and Hai Zhao. 2023b. Igniting Language Intelligence: The Hitchhiker's Guide From Chain-of-Thought Reasoning to Language Agents.

Pei Zhou, Aman Madaan, Srividya Pranavi Potharaju, Aditya Gupta, Kevin R. McKee, Ari Holtzman, Jay Pujara, Xiang Ren, Swaroop Mishra, Aida Nematzadeh, Shyam Upadhyay, and Manaal Faruqui. 2023. How FaR Are Large Language Models From Agents with Theory-of-Mind?

Pei Zhou, Jay Pujara, Xiang Ren, Xinyun Chen, Heng-Tze Cheng, Quoc V. Le, Ed H. Chi, Denny Zhou, Swaroop Mishra, and Huaixiu Steven Zheng. 2024. Self-Discover: Large Language Models Self-Compose Reasoning Structures.

Caleb Ziems, Minzhi Li, Anthony Zhang, and Diyi Yang. 2022. Inducing Positive Perspectives with Text Reframing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3682–3700, Dublin, Ireland. Association for Computational Linguistics.

R. A. Zwaan and G. A. Radvansky. 1998. Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.