Supplementary Materials

"From Lexicons to Large Language Models: A Holistic Evaluation of Psychometric Text Analysis in Social Science Research"

Appendix A: Case 1- Resilience in Governors' Tweets

In this case, we borrow from (Mousavi and Gu 2024) to measure the extent to which a given tweet contains resilience content. Resilience is defined as a "process of reintegrating from disruptions" (Richardson 2002, p. 309) or "adapting to a new normal after disruptions" (Buzzanell 2010) in life. Subsequently, Mousavi and Gu (2024) define resilience content as content that emphasizes on being able to adapt to change and content that emphasizes on the tendency to bounce back after hardship.

Mousavi and Gu kindly shared their data along with their resilience dictionary with us. We used the data for training and finetuning our models. We also reserved a portion of their data as held-out for testing. Given that the detailed description of the data and procedure for data labeling and training/ fine-tuning models are outlined in their online appendix (please see https://pubsonline.informs.org/doi/abs/10.1287/isre.2021.0599), we do not repeat these details here.

To create our custom-built model, we utilized the Long Short-Term Memory (LSTM) model, a type of deep learning model well-suited for working with text data. The details regarding how we trained this LSTM model are described in Appendix D. For FLMs, we fine-tuned several pre-trained models (please see appendix E).

We used three prominent LLMs, namely GPT4o, Llama 3.1, and Mistral to gauge resilience content within our samples. GPT4o, introduced by OpenAI in May 2024, stands as the pinnacle of state-of-the-art (SOTA) commercial LLMs. Llama 3.1, unveiled by Meta in July 2024, represents a suite of LLMs recognized as one of the best open-source LLM families, a title also attributed to Mistral v0.3, released by Mistral AI in May 2024. Within the Llama 3.1 family, we opted for 8B and 70B Instruct versions. Within Mistral v0.3, we used 7B Instruct version.

Given that the performance of these LLMs depends on the prompt, we examined a variety of prompting strategies reported in Appendix F.

Appendix B: Case 2- Empathy in Corporate Earnings Calls

Empathy has been defined and conceptualized in numerous ways over the past century, leading to inconsistencies in measurement and application (Clark et al. 2019, Cuff et al. 2016). While early definitions characterized empathy as the imaginative ability to take the perspective of another (Wispé 1986), subsequent research established empathy as a multidimensional construct encompassing both cognitive (i.e., understanding others' internal states) and affective components (i.e., feeling congruent emotions with others). While a consistent definition remains elusive in the management literature to date, an often-used description in the literature characterizes empathy as an other-oriented emotional response elicited by and congruent with the perceived welfare of someone else (Batson et al. 2002). Empathy ultimately involves perspective-taking, emotional contagion, and empathetic concern toward others in need (König et al. 2020, Salovey and Mayer 1990). Specific to our investigation of CEO speech, empathic communication is often defined as intentional behavior that demonstrates cognitive and/or affective empathy to others (Clark et al. 2019).

The multidimensional nature of empathy poses persistent challenges for measurement. Self-report assessments of perceived empathic tendencies face limitations such as social desirability biases (Chan 2010). Coding linguistic and nonverbal behavioral expressions of empathy involves highly subjective judgments of emotional content and intent (Zaki and Ochsner 2012). Empathy dictionaries can efficiently score large text samples but fail to adequately capture contextual nuances critical for assessing such a complex construct (Cuff et al. 2016). For example, leaders balance expressions of empathy with objectivity in communications to stakeholders. Simple word counts cannot determine if expressed emotions represent empathic perspective-taking versus personal distress. Studies have consistently underscored the relative efficacy of machine learning in this domain,

revealing its superior predictive validity when compared to conventional self-reporting and third-party evaluation methods of psychological constructs (Kern et al. 2016, Kosinski et al. 2016).

However, realizing the potential of AI-enabled empathy detection requires a clear conceptual foundation. Ambiguity in empathy definitions and uncertainty regarding its manifestation in leader communications have constrained previous automated analysis efforts. Synthesizing empathy's multidimensional nature with contextual insights will be imperative for developing AI beyond lexicon methods to uncover nuanced, empathetic expressions in leader-follower interactions.

1. Corporate Earnings Data

We initiated our data collection process by collecting the list of S&P 500 companies from Wikipedia.[1] Subsequently, for each of these companies, we made API calls to Finnhub.io's Stock API. Within each API call, our request encompassed all accessible corporate earnings transcripts. Our dataset spans a wide timeframe, ranging from the earliest earnings call recorded on October 30th, 2005, to the most recent on May 2nd, 2021. It is worth noting that, despite our efforts, we encountered data unavailability for certain S&P 500 firms. Specifically, we were unable to obtain data for the following companies: Berkshire Hathaway Inc. (BRK-B), International of Washington, Inc. (EXPD), NVR, Inc. (NVR), T. Rowe Price Group, Inc. (TROW), and Viatris Inc. (VTRS). In summary, our data collection efforts yielded a total of 24,906 earnings transcripts, covering 497 firms. This equates to an average of approximately 50 earnings transcripts per firm.

2. Data for Manual Coding

We began by extracting each sentence from all earnings calls in our dataset. From this extensive collection, we randomly selected a sample of 3,299 sentences, which were then subjected to evaluation by two independent human annotators. These annotators were tasked with reading each sentence and rating it based on the degree to which they perceived it to be related to empathy. For their efforts, the annotators received compensation at a rate of $10 per hour and collectively dedicated 30 hours to this project.

We provided them with a clear definition of empathetic language, emphasizing its focus on three key aspects: (1) the perception of others as in need, (2) the adoption of others' perspectives, and (3) the valuing of others' welfare. Using a five-point Likert scale, they were asked to assess each sentence's alignment with empathetic language, with response options ranging from "Not at all" to "Slightly," "Moderately," "Very," and "Extremely." Both evaluators diligently coded all 3,299 samples, and the inter-rater reliability score was 0.89.

As our goal was to employ this data for training and evaluating a deep learning model designed to automatically detect empathetic language, we transformed the five-point Likert scale into a binary scale. In this binary scale, a sentence was categorized as "Empathy" if both evaluators rated it as "Moderately," "Very," or "Extremely." Otherwise, it was labeled as "No Empathy." After applying this binary scale, our dataset was narrowed down to sentences that both evaluators agreed on, either as Empathy or Non-Empathy, resulting in a total of 2,755 sentences. Among these, only 110 sentences were initially labeled as Empathy. Further examination of these 110 samples revealed that sentences coded as "Moderately" by both evaluators did not consistently contain empathetic language. Consequently, we introduced an additional criterion: a sentence would be considered to contain empathetic language if at least one of the two evaluators rated it as "Very" or "Extremely," while the other evaluator gave it a rating of at least "Moderately." This refinement led to the identification of 50 *Empathy* samples, constituting 1.8% of the rated sentences, alongside 2,705 *Non-Empathy* samples.

3. Detecting Empathetic Language in Text Data

3.1. Using Lexicons to Detect Empathetic Language

Previous studies have used lexicon-based approaches to automatically detect empathetic language in text data (Sedoc et al. 2019, Sergent and Stajkovic 2020). In the lexicon-based approach, keywords related to empathetic language are detected in text documents. Then, a weighted score (based on the known strength of the relationship between the keyword and empathetic language) is calculated. In what is claimed to be the first lexicon for detecting empathy, Sedoc and colleagues train a model to predict document-level empathy in a regular supervised set-up and then "invert" the resulting model to derive word ratings (the lexicon). The authors

[1] https://en.wikipedia.org/wiki/List_of_S%26P_500_companies

use a Mixed-Level Feed Forward Network (MLFNN) on the document level using a neural Bag of Words (BoW) approach with an external, pre-trained embedding model. The data for training the document-level MLFNN is called EmoBank, which is a large-scale corpus manually annotated with emotion according to the psychological Valence-Arousal-Dominance scheme (Buechel and Hahn 2017). After obtaining the word ratings and creating the lexicon, Sedoc and colleagues evaluate their lexicon by comparing its performance with a variety of other lexicons and supervised models.

The second lexicon developed to measure empathetic language in text data is introduced in (Sergent and Stajkovic 2020). Sergent and Stajkovic combined several lexicons from Linguistic Inquiry and Word Count (LIWC) to construct a lexicon for measuring empathetic language. This lexicon is based on the following dictionaries in LIWC: feelings, death, money, work, and confidence (reward). After constructing the empathy lexicon, the authors use it to detect empathetic language in U.S. Governors' statements during the COVID-19 pandemic.

3.2. Using Custom-built & Fine-tuned NLP Models to Detect Empathetic Language

Our full dataset, labeled by independent evaluators, contains 2,755 sentences from corporate earnings calls. We partitioned this data into a 2,000-sentence training set and a 755-sentence validation set. The training set was used to build the CBMs and FLMs. The validation set initially consisted of 12 *Empathy* and 743 *Non-empathy* sentences. We ultimately removed 28 of the *Non-empathy* samples because at least one of the LLMs (in Table 2 or 9) failed to return valid probability score (e.g., providing a text justification instead). This resulted in a final test set of 727 sentences for our evaluation.

To address the extreme label imbalance within the training set, where only 38 sentences were initially labeled as empathetic language among the sentences in the training data, we used the back-translation approach that we used in Mousavi & Gu (2024) to synthetically generate additional positive (*Empathy*) sentences. We leveraged Google Translate's API for this purpose, translating the Empathy-labeled sentences into various supported languages, such as Spanish, German, French, and Arabic. Subsequently, we translated these sentences back to English using the same API. By employing this approach, we augmented our dataset with 400 *Empathy* samples, comprising the 38 original *Empathy* samples and an additional 362 synthetically generated ones. We maintained the same number of samples from the *Non-empathy* collection to ensure balance, resulting in a total of 400 *Empathy* and 400 *Non-empathy* samples for training our NLP models.

In our first NLP model, we utilized an LSTM model. The details regarding how we trained this LSTM model are described in Appendix D. For the second NLP model, we fine-tuned a variety of pre-trained models (please see Appendix E for details and performance results).

3.3. Using LLMs to Detect Empathetic Language

Similar to case 1, we used GPT4o, Llama 3.1 and Mistral v0.3 to label our test samples in case 2. As noted in Appendix F, we tested a variety of prompting strategies.

Appendix C: Case 3- Polarity in Financial News Headlines

Analyzing the polarity of text is a common task that has been used in a variety of studies (Tetlock 2007, Tetlock et al. 2008). Hence, to further examine our findings in another common data labeling case, we decided to replicate our approaches with analyzing polarity in financial news headlines.

1. News Headlines Data

We sourced our labeled dataset from Malo et al. (Malo et al. 2013). This dataset consists of roughly 5,000 phrases or sentences extracted from financial news and company press releases. A panel of 16 annotators, all with a business education background, categorized each phrase or sentence as positive, negative, or neutral (Malo et al. 2013). Given that the lexicon (Loughran and McDonald 2011) that we use in this case only detects positive and negative polarities, we excluded the neutral samples from our data. This left us with 1,967 labeled samples. Of these, we used 1,500 samples chosen at random for training and the remaining 467 (315 *positive* and 152 *negative*) samples for testing our NLP methods. We ultimately removed one of the *positive* samples because at least one of the LLMs (in Table 2) failed to return valid probability score (e.g., providing a text justification instead). This resulted in a final test set of 466 samples for our evaluation.

2. Detecting Polarity in Financial News Headlines

Similar to cases 1 and 2, we use a lexicon-based method, an LSTM model, a fine-tuned BERT model (e.g., FinBERT), Llama, Mistral, and GPT4o to measure the polarity in text.

2.1. Using a Lexicon to Detect Polarity

For the lexicon-based approach, we use a widely used lexicon developed by Loughran and McDonald (2011). This lexicon, which we will refer to as the L&M lexicon, is built upon the 4.0 release of 2of12inf—a standard lexicon that emphasizes common words. The 2of12inf omits abbreviations, acronyms, British English terms, hyphenated words, names, and phrases but includes word inflections.

The L&M lexicon enhanced the base lexicon by incorporating terms commonly found in 10-K documents and earnings calls but absent from the original 2of12inf list. This enrichment was informed by a thorough analysis of the EDGAR 10-K archive and earnings calls from CapIQ. The lexicon categorizes words into negative, positive, uncertainty, litigious, strong modal, weak modal, and constraining, as per Loughran and McDonald (2011). Our interest was primarily in the polarity; thus, we utilized only the positive and negative word sets. Moreover, we restricted our use to words present in the lexicon's latest version, excluding terms that were removed in 2020.

Before using the L&M lexicon on our dataset, we pre-processed both the lexicon text and our labeled sample. This involved converting text to lowercase, applying lemmatization (using the WordNet lemmatizer from NLTK), and removing non-alphabetic characters. Next, we searched our test samples for occurrences of positive and negative words from the L&M lexicon. By dividing the count of positive/negative words by the total words in each sample, we derived a *positive* and a *negative* measure. We first calculated the AUC score for each one separately. The AUC associated with *positive* measure was 61.24%, whereas the AUC for *negative* (inversed) was 67.65%. We then combined *positive* and *negative* measures by deducting *negative* from *positive*. This gave us the best AUC on the test data (71.64%).

2.2. Using Custom-built, Fine-tuned NLP Models, & LLMs to Detect Polarity

We used the 1,500 labeled samples described earlier to train an LSTM model from scratch. The details regarding training the LSTM model are described in Appendix D. As reported in Appendix E, we fine-tuned a variety of pre-trained models for this task, similar to the previous two. For LLMs, we used the same models and prompting strategies as reported in the manuscript and Appendix F.

Appendix D: Training the Long Short-Term Memory (LSTM) Model

Given that the predictive performance of our LSTM model depends on its architecture and the hyper-parameter values that we use, we decided to deploy the Python package "KerasTuner"[2], an open-source library designed for hyperparameter tuning in deep learning models (specifically tailored for the Keras deep learning framework), to try a variety of architectures (number of LSTM, dense, and drop-out layers) as well as parameters (the learning rate, the activation on the dense layers, the size of the embedding layer, the drop-out ratio, …) to identify the best setting. KerasTuner's workflow consists of the following steps:

1. Define the search space and model-building function, specifying the hyperparameters and their respective ranges or choices.
2. Instantiate the tuner, selecting the desired search algorithm, objective function, and other tuning parameters, such as the maximum number of trials or the number of models to be evaluated in parallel.
3. Launch the search process using the tuner's search() method, which iteratively samples hyperparameter configurations, builds and trains the corresponding models, and evaluates their performance on the specified objective function.
4. Upon completion, retrieve the best hyperparameter configuration and the associated performance metric from the tuner.

Specifically, we used the following search space:

[2] Documentation available in https://keras.io/keras_tuner/.

1. The number of LSTM layers: 1 to 2
2. The number of units in each LSTM layer: 64, 128, 256
3. The number of Dense layers: 1 to 2
4. The number of units in each Dense layer: 64, 128, 256
5. The Dropout rate (ratio) after the last Dense layer and before the output layer: 0.3, 0.5, 0.7
6. The learning rate in the optimizer: 1e-2, 1e-3, 1e-4

In addition to the architecture and hyper-parameters of the deep learning model, we tried different values for vocabulary size. The vocabulary size determines the number of unique words (or tokens) the model will recognize and process. The vocabulary size can affect the model's performance, memory requirements, and computational complexity. Changing the vocabulary size can affect various aspects of the model: Reducing the vocabulary size may lead to faster training and inference times and lower memory requirements. However, this might also result in a loss of information if less frequent but meaningful words are excluded from the vocabulary. Increasing the vocabulary size can capture more information from the text data and potentially improve the model's performance. However, it also increases the memory requirements and computational complexity, which could lead to longer training and inference times. To determine the right value for vocabulary size, we tried 1,500, 2,000, 2,500, 3,000, 3,500, and 4,000 as the values and compared the AUC scores on the validation set. Given that the vocabulary size, the network architecture, and the hyper-parameter values can interact, we ran KerasTuner several times each time for each vocabulary size mentioned above. KerasTuner identified the architecture represented in Figure D and Table D for each one of the three cases in our study.

```
inputs = Input(name='inputs',shape=[max_len])
layer = Embedding(input_dim,output_dim,input_length)(inputs)
layer = LSTM(num_lstmu)(layer)
layer = Dense(num_denseu,name='FC1')(layer)
layer = Activation('relu')(layer)
layer = Dropout(drop_rate)(layer)
layer = Dense(1,name='out_layer')(layer)
layer = Activation('sigmoid')(layer)
model = Model(inputs=inputs,outputs=layer)
return model
```

Figure D. The Architecture of the LSTM Model

Table D. The Hyper-parameter Values for the LSTM Model Across Cases

Hyper-parameter	Value for Empathy	Value for Resilience	Value for Polarity
input_dim	2,500	2,000	4,000
output_dim	100	50	50
input_length	35	50	50
num_lstmu	128	64	64
num_denseu	256	256	256
drop_rate	0.3	0.5	0.5
learning_rate	0.01	0.01	0.01

Appendix E: Fine-tuning Pre-trained Models (FLMs)

In our approach, we employ an embedding layer for token, position, and token type embeddings. This layer feeds into a combination of attention and dense layers, culminating in an output layer with several pooling and dense layers. This setup lets us utilize BERT embeddings and fine-tune them with our data. For model optimization, we chose the AdamW optimizer (Adam with weight decay correction). We used a grid search to find reasonable values for learning rate, batch size, weight decay, and the number of training epochs. We tried "bert-base-uncased", "bert-large-uncased", "roberta-base", "roberta-large", "albert-base-v2", "albert-large-v2", "distilbert-base-uncased", and "distilroberta-base" pre-trained models and selected the one with the highest

AUC score for the resilience and empathy cases. By default, BERT base uses 512 as the value for the maximum sequence length. However, given that our samples are much shorter than that, we used a maximum length of 100 for the empathy case and 111 for the resilience case (obtained based on the same process described in Mousavi & Gu (2024)). Before using our BERT- based model, we used the BERT tokenizer to encode the text. This tokenizer uses sub-word-based vocabularies (e.g., WordPiece) to tokenize the text. Please refer to https://huggingface.co/docs/transformers/main_classes/tokenizer for more information about the tokenizer. Table E reports the highest performance achieved by fine-tuning each pre-trained model for case.

Table E. FLMs' Performance Across Cases

Model	AUC for Resilience	AUC for Empathy	AUC for Polarity
albert-base-v2	0.980857	**0.922104**	0.990129
albert-large-v2	0.896701	0.689919	0.542439
bert-base-uncased	0.984576	0.847344	0.997104
bert-large-uncased	0.796921	0.715985	0.995233
distilbert-base-uncased	0.970532	0.900664	0.996724
distilroberta-base	**0.985883**	0.857932	0.997280
roberta-base	0.971263	0.85367	**0.998976**
roberta-large	0.635028	0.748241	0.594999

Appendix F Prompting Techniques, Soft Prompting, & Fine-tuning LLMs

The five prompting strategies we evaluated were:

1. Few-Shot Prompting: This technique involves providing the model with a limited number of input-output examples to guide its responses. By presenting these exemplars, the model can infer the desired task and generate appropriate outputs. This approach was the original prompting strategy proposed by GPT-3 team (Brown et al. 2020).

2. Chain-of-Thought (CoT) Prompting: CoT prompting encourages the model to produce intermediate reasoning steps, thereby facilitating complex problem-solving. By decomposing tasks into sequential steps, the model's reasoning process becomes more transparent and effective. This method was introduced in (Wei et al. 2023).

3. Contrastive Chain-of-Thought Prompting: Building upon CoT, this strategy incorporates both valid and invalid reasoning demonstrations. By exposing the model to correct and incorrect examples, it learns to discern and emulate valid reasoning patterns. Proposed in (Chia et al. 2023).

4. Tree of Thoughts (ToT): This method extends the CoT framework by exploring multiple reasoning paths simultaneously, akin to a tree structure. By evaluating various branches of thought, the model can select the most promising line of reasoning. This technique was introduced by Yao et al. (2023).

5. Self-Refine: This strategy enables the model to iteratively refine its responses by self-critiquing and improving upon initial outputs. Through successive iterations, the model enhances the quality and accuracy of its answers. Madaan et al. (2023) introduced this technique in "Self-Refine: Iterative Refinement with Self-Feedback".

Per results reported in Table 7 in the manuscript, Contrastive CoT resulted in the best performance for Resilience and Empathy whereas ToT resulted in the best performance for Polarity. We used these prompting techniques in Tables 2, 3, and 4. Below, we present the Contrastive CoT prompt for Empathy annotation and ToT prompt we used for Polarity annotation.

Contrastive CoT for Empathy Annotation:
"""Act as an expert human coder who classifies sentences into high and low empathy. Learn by examining both correct and incorrect reasoning in the examples provided.
Example 1:

Sentence: "When Johnson & Johnson takes this patient-centric view of chronic and pervasive conditions like Diabetes or heart disease, it can deploy all its assets and medical knowledge to serve patients more effectively."
Correct Reasoning: This sentence emphasizes a patient-centric approach, highlighting empathy towards individuals with chronic conditions.
Classification: High empathy.
Incorrect Reasoning: This sentence mainly discusses the effectiveness of deploying assets, without focusing on a patient-centric view.
Incorrect Classification: Low empathy.
Example 2:
Sentence: "When Johnson & Johnson takes this view of chronic and pervasive conditions like Diabetes or heart disease, it can deploy all its assets and medical knowledge more effectively."
Correct Reasoning: This sentence lacks a patient-centric focus and does not convey empathy towards individuals with chronic conditions.
Classification: Low empathy.
Incorrect Reasoning: This sentence mentions patients, suggesting an implicit focus on their conditions, which indicates empathy.
Incorrect Classification: High empathy.
Now, determine the probability of high empathy for the following sentence. Consider both the correct and incorrect reasoning steps. Provide only the probability of high empathy as a float number between 0 and 1, rounded to two decimal places.
The sentence: {}"""

ToT for Polarity Annotation:
"""Act as an expert human coder who can classify sentences into positive or negative sentiment. For each sentence, generate multiple reasoning paths (thoughts) that could lead to different sentiment classifications. Evaluate each path, considering language indicating positivity, negativity, or neutrality, and the tone of the statement. After exploring these paths, determine the probability of the sentence exhibiting positive sentiment. Provide only the probability as a float number between 0 and 1, rounded to two decimal places.
Example Sentence:
"The company's stock price saw a significant rise compared to the last quarter."
Reasoning Paths:
Path 1:
Observation: The sentence mentions a "significant rise" in stock price.
Evaluation: The language suggests positive sentiment due to financial growth.
Conclusion: Positive sentiment.
Path 2:
Observation: Compares stock performance to the last quarter positively.
Evaluation: Indicates progress, aligning with positive sentiment.
Conclusion: Positive sentiment.
Self-Evaluation:
Given the consistent positive language and tone, the sentence likely exhibits positive sentiment.
Probability of Positive Sentiment: 0.95
New Sentence for Classification:
"The company's stock price saw a significant drop compared to the last quarter."
Reasoning Paths:
Path 1:
Observation: The sentence mentions a "significant drop" in stock price.
Evaluation: The language suggests negative sentiment due to financial decline.
Conclusion: Negative sentiment.
Path 2:

Observation: Indicates a comparison that highlights underperformance.
Evaluation: Suggests negativity in the context of financial outcomes.
Conclusion: Negative sentiment.
Self-Evaluation:
The language and focus on a negative trend indicate the sentence likely exhibits negative sentiment.
Probability of Positive Sentiment: 0.10
Now, determine the probability of positive sentiment for the following sentence. Consider different reasoning paths. Provide only the probability of positive sentiment as a float number between 0 and 1, rounded to two decimal places.
The sentence: {}"""

For cognitive–affective prompting, we utilized the best traditional prompting strategy for each case and incorporated a statement to highlight cognitive abilities and/or emotional intelligence. Below, we provide the cognitive–affective prompt used for Empathy annotation, with modifications to the traditional prompt highlighted in bold.

"""Act as an expert human coder **who possesses exceptional emotional intelligence and superior cognitive abilities, enabling you to** classify sentences into high and low empathy. Learn by examining both correct and incorrect reasoning in the examples provided.
Example 1:
Sentence: "When Johnson & Johnson takes this patient-centric view of chronic and pervasive conditions like Diabetes or heart disease, it can deploy all its assets and medical knowledge to serve patients more effectively."
Correct Reasoning: This sentence emphasizes a patient-centric approach, highlighting empathy towards individuals with chronic conditions.
Classification: High empathy.
Incorrect Reasoning: This sentence mainly discusses the effectiveness of deploying assets, without focusing on a patient-centric view.
Incorrect Classification: Low empathy.
Example 2:
Sentence: "When Johnson & Johnson takes this view of chronic and pervasive conditions like Diabetes or heart disease, it can deploy all its assets and medical knowledge more effectively."
Correct Reasoning: This sentence lacks a patient-centric focus and does not convey empathy towards individuals with chronic conditions.
Classification: Low empathy.
Incorrect Reasoning: This sentence mentions patients, suggesting an implicit focus on their conditions, which indicates empathy.
Incorrect Classification: High empathy.
Now, determine the probability of high empathy for the following sentence. Consider both the correct and incorrect reasoning steps. Provide only the probability of high empathy as a float number between 0 and 1, rounded to two decimal places.
The sentence: {}"""

For soft-prompting, we used parameter-efficient fine-tuning (PEFT) technique that adapts LLMs to specific tasks by introducing trainable prompt tokens into the input sequence. Unlike traditional fine-tuning methods that adjust all model parameters, prompt tuning keeps the original model weights frozen and optimizes only the additional prompt embeddings. This approach reduces computational costs and storage requirements while maintaining or enhancing performance on downstream tasks. We followed the instructions outlined in https://github.com/huggingface/peft for soft prompting.

For fine-tuning, we used Low-Rank Adaptation (LoRA), which introduces trainable low-rank matrices into each layer of a pre-trained model, allowing for task-specific adaptation without updating the full set of model parameters. We adapted from https://kickitlikeshika.github.io/2024/07/24/how-to-fine-tune-llama-3-models-with-LoRA.html to finetune the open-source LLMs in our study. For fine-tuning GPT4o, we used OpenAI's API, as described in the manuscript.

Appendix G: Fairness Analysis

We used the data compiled by Abbasi et al. (2021) to conduct this analysis. The dataset preparation involved a comprehensive process focusing on psychometric dimensions such as trust, anxiety, literacy, and numeracy within the health context. The dataset combines user-generated text and survey-based psychometric measures from 8,502 respondents, along with their self-reported demographic information.

The dataset's construction began with a literature review in behavioral health, which identified relevant psychometric dimensions. This review informed the selection of survey-based items to operationalize these dimensions into measurable constructs. For each dimension, appropriate survey items were developed or adapted from existing scales validated through exploratory and confirmatory factor analysis. These items are designed to capture aspects of the respondents' health literacy, numeracy, trust in physicians, and anxiety regarding visiting the doctor. To gather user-generated text, the study employed an iterative trial-and-error approach to develop prompts that effectively elicited relevant text responses from participants. This involved refining the placement and wording of text-response prompts to align closely with the survey items. Respondents were then primed with these survey items before being asked to provide text responses that reflected their thoughts and feelings relevant to each psychometric dimension.

The dataset also includes demographic data for each participant, including race and gender, which allows for the analysis of biases and the fairness of text classification methods used within the study. The combination of textual, survey, and demographic data provides a rich resource for examining the relationships between user-generated text and standardized psychometric measures, facilitating research on fairness and bias in natural language processing (NLP) applications.

To create the predicted labels, we applied each one of our psychometric NLP methods (i.e., lexicons, custom-built, FLMs, and SOTA LLMs) to measure each one of the four psychometric attributes (i.e., health literacy, numeracy, trust in physicians, and anxiety). As noted before, we used DI to measure fairness with respect to gender and race. We defined our privileged/non-privileged groups exactly the same way Abbasi et al. (2021) we defined these groups, based on these two factors: gender ("male" vs. "non-male") and race ("white" vs. "non-white"). We reserved 20% of the samples that passed the survey's quality checks for comparing the four NLP methods and used the rest to train/ fine-tune models when needed.

In applying the lexicon-based method, we first used the following lexicons:

1- For health literacy: LIWC's (Pennebaker et al. 2015) analytic, sixteen readability scores populated by Python package "textstat."[3]
2- For numeracy: LIWC's analytic, number, and quant
3- For trust in physician: LIWC's clout and authentic
4- For anxiety when visiting a physician: LIWC's anxiety, anger, sad, and negative, and the anxiety lexicon developed by (Rheault 2016).

After implementing various lexicons on the dataset, none yielded an AUC score higher than 0.55, indicating that lexicon-based methods are not effective for assessing the four targeted attributes. Consequently, we shifted to a hybrid approach that integrates the LIWC with supervised machine learning. In this approach, by processing each text sample with LIWC version 15, we captured over 90 different textual attributes. These attributes were then utilized in an XGBoost classifier trained on 80% of the labeled data. To improve the predictive performance of the models, we used the Python package Optuna[4] for hyperparameter optimization.

[3] https://pypi.org/project/textstat/

[4] https://optuna.org

This improved approach achieved AUC scores between 0.667 and 0.702, aligning with the results reported by Abbasi et al. (2021).

For the custom-built model, we employed an LSTM model as detailed in Appendix D. For the FLM method, we utilized BERT-based methods, as described in Appendix E. For the SOTA LLM method, we applied Contrastive CoT as it produced the most accurate results in the previous cases:

"""In a study, participants were asked to what extent they feel anxious when they see their doctor. Then, they were asked to "describe what makes them feel most anxious or worried when visiting the doctor's office."
Now, act as an expert annotator who can classify documents based on participants' reasons for feeling anxious when seeing a doctor.
Learn by examining both correct and incorrect reasoning in the examples provided.
Example 1:
Sentence: "I get nervous that they are going to tell me that something is wrong or that I need medication to better my health."
Correct Reasoning: This sentence expresses fear and concern about receiving bad news or needing medication, which strongly indicates high anxiety.
Classification: High anxiety.
Incorrect Reasoning: This sentence only describes a possible situation and does not explicitly mention the participant's feelings, suggesting low anxiety.
Incorrect Classification: Low anxiety.
Example 2:
Sentence: "I never get nervous when they tell me that something is wrong or that I need medication to better my health."
Correct Reasoning: This sentence explicitly states the absence of nervousness, which clearly indicates low anxiety.
Classification: Low anxiety.
Incorrect Reasoning: The mention of medical issues implies a potential for worry, which might suggest high anxiety.
Incorrect Classification: High anxiety.
Now, determine the probability of high anxiety for the following sentence. Consider both the correct and incorrect reasoning steps. Provide only the probability of high anxiety as a float number between 0 and 1, rounded to two decimal places.
The sentence: {}"""

Appendix H: Instrument Validity and Qualitative Analysis

H1. Convergent and Divergent Validity of Key Instruments

We evaluated four theoretically distinct constructs—Cognitive Ability, Emotional Intelligence, Positive Mood (Pos_Mood), and Negative Mood (Neg_Mood)—via separate single-factor maximum-likelihood models fit to z-standardized item sets. For each construct we report internal consistency (Cronbach's α), McDonald's ω, composite reliability (CR), average variance extracted (AVE), corrected item–total correlations, the Kaiser–Meyer–Olkin (KMO) index, and Bartlett's test of sphericity. Discriminant validity among the four factors was assessed using the Fornell–Larcker criterion—comparing the square root of AVE (diagonal) to inter-construct latent correlations (off-diagonals)—and the heterotrait–monotrait ratio (HTMT). Factor loadings were sign-aligned for interpretability. Composite scores were created per our preregistered rules: Cognitive Ability as the sum of its eight items and the other three constructs as item means.

Cognitive Ability demonstrated excellent reliability ($\alpha = .902$, $\omega = .899$, CR = .899), solid convergent validity (AVE = .530), KMO = .884, and a significant Bartlett test ($\chi^2 = 3{,}218.20$, $p < .001$); loadings ranged .57–.84 with corrected item–total r = .62–.75. Emotional Intelligence showed excellent reliability ($\alpha = .929$, $\omega = .929$, CR = .929) with AVE = .453 (slightly below the .50 convention), KMO = .918, Bartlett $\chi^2 = 7{,}159.99$ ($p < .001$), and loadings .48–.79 (item–total r = .49–.74). Pos_Mood exhibited strong convergent evidence ($\alpha = .925$,

ω = .925, CR = .925, AVE = .556; KMO = .927; Bartlett χ^2 = 3,936.51, $p < .001$) with loadings .55–.85 (item–total r = .56–.80). Neg_Mood showed similarly strong properties (α = .921, ω = .922, CR = .922, AVE = .543; KMO = .911; Bartlett χ^2 = 3,747.94, $p < .001$) with loadings .58–.81 (item–total r = .56–.76). Discriminant validity was supported: the square roots of AVE for all constructs (Cognitive Ability = .728; Emotional Intelligence = .673; Pos_Mood = .746; Neg_Mood = .737) exceeded their inter-construct correlations (largest |r| = .583 between Emotional Intelligence and Pos_Mood), and the maximum HTMT was .633—well below conservative thresholds (≤ .85). The latent correlation pattern aligned with theory, with Pos_Mood inversely related to Neg_Mood (r = −.206), Cognitive Ability modestly positively related to Emotional Intelligence (r = .297), and Cognitive Ability moderately inversely related to Neg_Mood (r = −.499). Overall, three constructs (Cognitive Ability, Pos_Mood, Neg_Mood) met conventional convergent benchmarks, Emotional Intelligence was highly reliable with marginal AVE, and discriminant validity was clearly established across all four domains.

H.2 Analysis of Participants' Highlighted Tokens

During our data collection, we asked each participant to highlight segments of the text samples that contributed to their decision to label the samples. Below, we present the results of analyzing these highlighted portions of text in our samples. For each construct, we divide the data into four groups based on *expert* and *participant* labels. For example, in Figure H1, the top left plot (expert=1 and participant=1) displays the most frequent words highlighted by participants in cases where they correctly labeled a sample as high resilience. In contrast, the top right plot (expert=1 and participant=0) displays the most frequent words highlighted by participants in cases where they incorrectly labeled a sample as low resilience.

It is worth noting that the frequencies are all normalized by dividing the count of each word in each group (e.g., expert = 1, participant = 1) by the total number of words in the group, ensuring that frequencies are comparable across groups of different sizes. In addition, we removed stop-words and words with fewer than 3 characters. We also stemmed and lowercased all the words in the data.

Resilience:

This section investigates how participants differentiated samples conveying high vs. low resilience, focusing on the specific words they highlighted. Figure H1 displays the top 20 most frequently highlighted words for each group:

1. Top-Left Plot (Expert=1, Participant=1): Correct High-Resilience Annotations

The words here—such as pandem, vaccin, continu, help, posit, recoveri, and move—often point to forward-looking or hopeful sentiments (e.g., recovery, continuation, keeping on track). Participants viewed these terms as indicative of high resilience, reflecting themes of adaptation and optimism amidst pandemic-related challenges.

2. Top-Right Plot (Expert=1, Participant=0): Missed High Resilience

In these texts, experts detected high resilience, but participants labeled them as low. Frequently highlighted words—vaccin, continu, keep, posit, forward, move, help, recoveri—reflect clear forward movement, positivity, or signs of adaptation. Participants may have overlooked these underlying resilience themes, possibly because many references (e.g., vaccin, pandem) also appeared in non-resilience contexts, leading to confusion.

3. Bottom-Left Plot (Expert=0, Participant=1): Incorrectly Labeled as High Resilience

Here, participants labeled texts as high resilience despite experts considering them low resilience. Common highlights include test, covid, help, hard, and mask. While certain words (e.g., help, posit) might suggest coping, the experts likely identified insufficient evidence of genuine resilience in the broader message context. Participants may have overinterpreted pandemic-related or neutral references (such as vaccin or work) as resilient cues.

4. Bottom-Right Plot (Expert=0, Participant=0): Correct Low-Resilience Annotations

Words highlighted in this group (e.g., test, posit, case, death, increas, hospit) commonly indicate neutral or negative pandemic facts rather than personal or communal resilience. Both participants and experts agreed these messages lacked elements of proactive coping or forward thinking.

These results highlight how participants gravitated to certain pandemic-related words—such as vaccin, test, covid, and posit—that can appear in both high- and low-resilience contexts. When participants and experts agreed on high resilience, the emphasis was on forward momentum (continu, recoveri, move, help, posit). However, participants sometimes misread these words in less clearly resilient contexts (leading to overestimation of resilience) or overlooked them where genuine resilience was present (leading to underestimation).

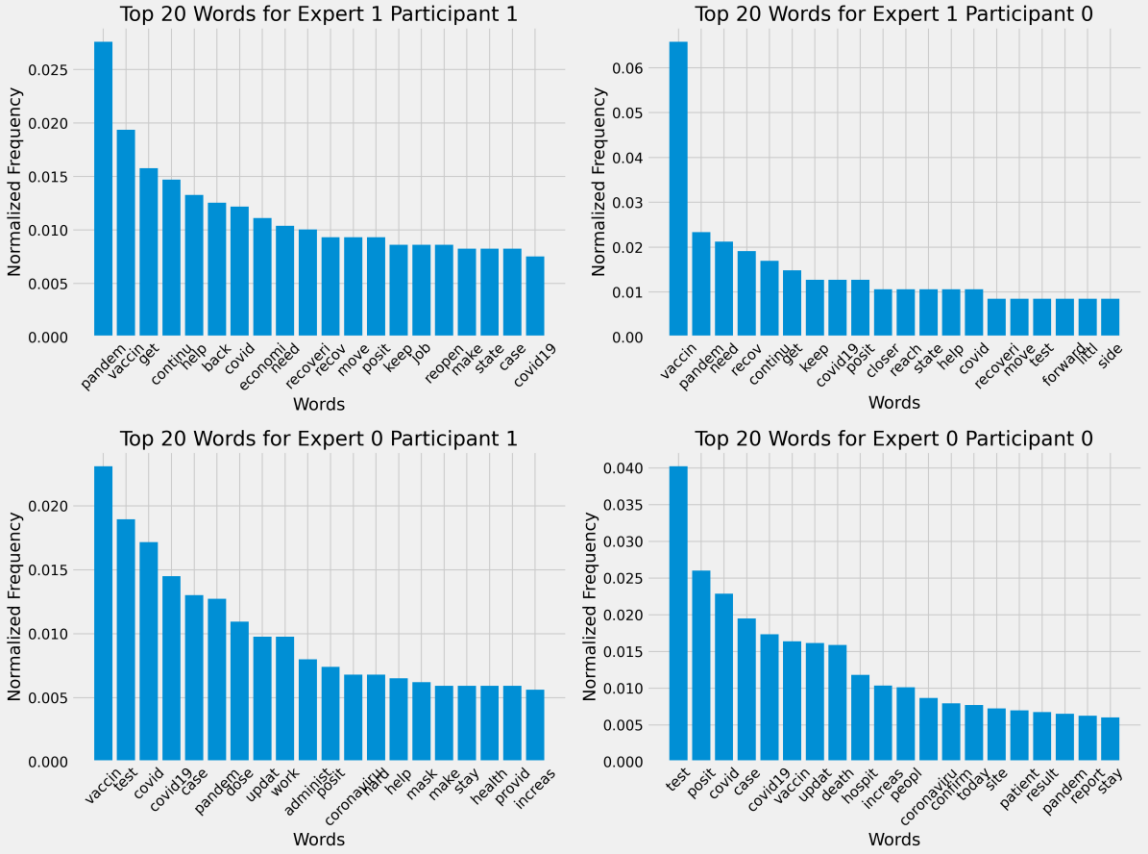

Figure H1. Normalized Word Frequencies for Each Group for Resilience Annotation

Empathy:

This section explores how participants identified text messages conveying high vs. low empathy. Similar to the resilience case, Figure H2 displays the top word frequencies divided based on the agreement/ disagreement between expert and participant annotations:

1. Top-Left Plot (Expert=1, Participant=1): Correct High-Empathy Annotations

The words in this category include benefit, custom, famili, parent, focus, enjoy, work, provid, team, partner, clean, full, fun, friendli, experi, gener, ensur, maximum, sympathet, and allergi. Many of these terms reflect supportive, considerate, or personable elements, such as famili, friendli, sympathet, and ensur, which naturally indicate empathy. Participants correctly deemed these cues indicative of genuine concern, care, or understanding.

2. Top-Right Plot (Expert=1, Participant=0): Missed High Empathy

In contrast, the experts recognized empathy in these texts, while participants labeled them as low empathy. Frequently highlighted words include custom, take, enjoy, care, congratul, commun, relev, proposit, ensur,

work, delight, sympathet, parent, famili, full, fun, friendli, connect, confid, and right. Critically empathetic words like care, sympathet, and friendli appear here, suggesting participants may have overlooked or minimized cues of genuine concern and warmth. This gap highlights instances where the context surrounding these terms could be key to recognizing empathy.

3. Bottom-Left Plot (Expert=0, Participant=1): Incorrectly Labeled as High Empathy

Despite expert evaluations of low empathy, participants highlighted words such as happi, like, team, sorri, enjoy, see, hurt, peopl, think, make, worri, thank, help, work, custom, would, delight, realli, group, and encourag. Some of these words (sorri, help, thank) can suggest concern or kindness, but they may not signal deeper empathetic engagement in context. The experts likely determined that these samples lacked authentic empathy despite containing superficially "positive" or polite language. Participants, however, may have overinterpreted such positivity as empathetic intent.

4. Bottom-Right Plot (Expert=0, Participant=0): Correct Low-Empathy Annotations

Both participants and experts agreed these samples lacked substantive empathetic cues. The top words in this group—like, happi, team, sorri, growth, year, market, see, think, enjoy, would, work, busi, continu, strong, hurt, sale, realli, help, benefit—span neutral, business-oriented, or superficially positive language that does not necessarily convey empathy. While terms like happi, sorri, or help can have empathetic undertones, the broader context likely did not support a genuine empathic stance.

These findings underscore how participants' attention to seemingly "positive" or "caring" words sometimes led them to over-identify or miss empathy in a sample. Words like care, friendli, or sympathet were strong indicators of empathy when used meaningfully in context (correctly labeled as high empathy), yet participants occasionally overlooked them (missed empathy) or mistook generic polite language (happi, team, thank) as empathy when it did not reflect authentic concern.

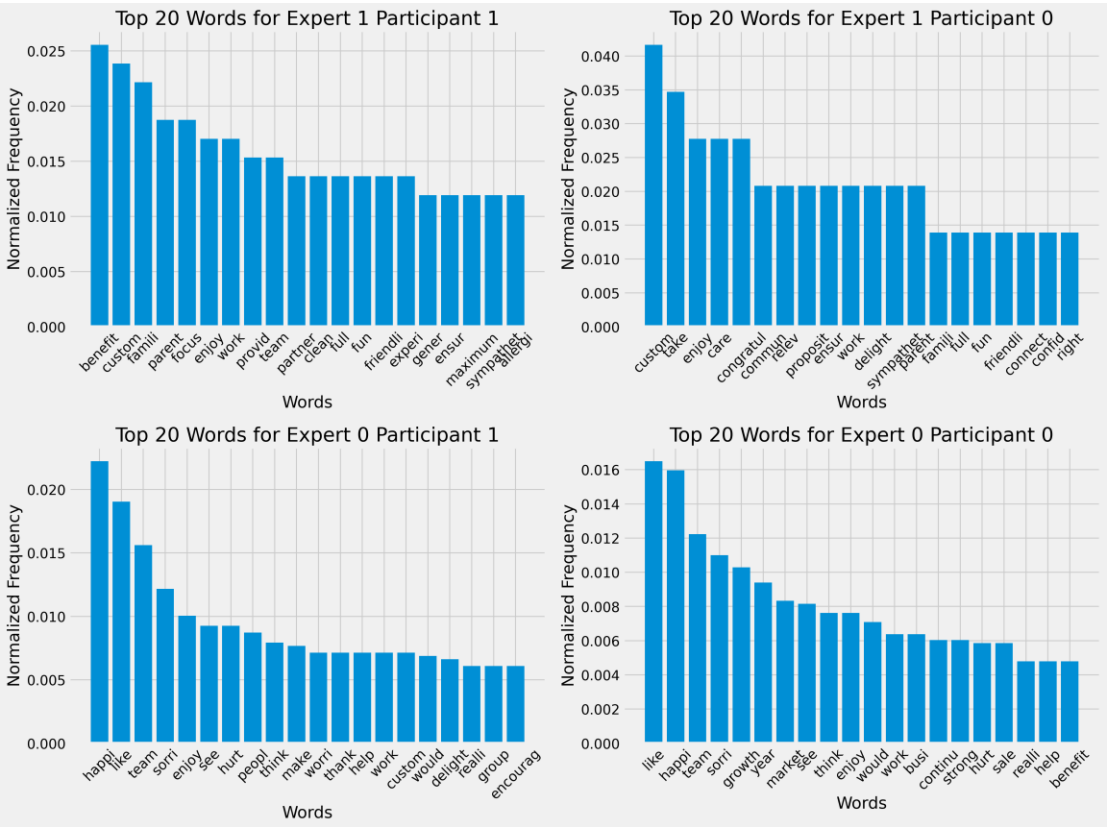

Figure H2. Normalized Word Frequencies for Each Group for Empathy Annotation

Our investigation into the nuanced interplay between emotional intelligence and the accuracy of empathy annotations has revealed unexpected discrepancies, necessitating a deeper analysis of the data. Specifically, we scrutinized cases where expert evaluators labeled statements as "low empathy," yet a significant majority of participants (≥80%) consistently annotated them as "high empathy." Below, we present representative examples, accompanied by a detailed synthesis of potential misclassification drivers:

1. **"I think as you can imagine, the conversation is really on supply and most of our conversations with our customers are all very emotional and all related to supply."**
    - *Participant interpretation:* The use of "emotional" was likely interpreted as indicative of concern for customer needs.
    - *Expert assessment:* The statement primarily focuses on operational supply issues and lacks substantive markers of empathy, such as perspective-taking, recognition of others' needs, or prioritization of welfare.
2. **"The whole idea behind is, we don't want our customers using third-party applications, particularly in what potentially is insecure -- in an unsecure fashion."**
    - *Participant interpretation:* Participants may have perceived the emphasis on protecting customers as prioritizing their welfare.
    - *Expert assessment:* The primary concern centers on organizational security rather than empathy, as defined within our framework.
3. **"I'm sorry, so the call center cost is in there, which is internal as well as we use some third parties for call center and then the third-party costs associated with the product that we're delivering to people for free."**
    - *Participant interpretation:* The apology and reference to a free product may have been construed as concern for others.
    - *Expert assessment:* The language is transactional and lacks evidence of perspective-taking or altruistic intent.
4. **"And I want to sincerely thank the Target team for their tireless effort to help our guests recover from the data breach."**
    - *Participant interpretation:* Expressions of gratitude and references to "helping guests recover" may evoke an impression of empathy.
    - *Expert assessment:* The statement reflects corporate responsibility rather than genuine perspective-taking or concern for individual welfare.
5. **"Sandy, when investors call us and they worry about EMEA, it seems like they worry about Hydraulics and given sort of what happened in that 2008 and 2009 period."**
    - *Participant interpretation:* Acknowledging investors' concerns might be interpreted as perspective-taking.
    - *Expert assessment:* The focus is on addressing business anxieties rather than demonstrating concern for others' needs or well-being.
6. **"In closing, I want to thank my amazing team for strategic thinking and tireless execution."**
    - *Participant interpretation:* Expressions of gratitude resonate emotionally.
    - *Expert assessment:* While appreciative, the statement lacks focus on adopting others' perspectives or addressing their needs.

Our findings reveal an intriguing paradox: individuals with high emotional intelligence, despite their heightened sensitivity to affective language, may overinterpret emotional cues—such as gratitude, apologies, or mentions of concern—as signals of empathy. This predisposition may result in systematic errors when discerning between surface-level emotional resonance and the deeper, analytical markers of substantive empathy, such as perspective-taking, recognizing others' needs, and valuing their welfare.

This phenomenon underscores the importance of rigorously defining and operationalizing empathy in annotation frameworks, particularly for tasks requiring nuanced judgment. Experts, for instance, employ a structured and principled approach, emphasizing key elements such as perspective-taking, acknowledgment of others' needs, and altruistic intent. Consider the expert-labeled example:

- *"While our team was focused on ensuring the safety of their own families, they also worked tirelessly to reopen stores quickly and move needed essentials from other parts of the country in support of our guests."*

This statement highlights selfless actions addressing others' needs during a crisis. Similarly:

- *"At the meeting, we respectfully announced a new breakthrough partner benefit, critical illness insurance to parents of our partners, a benefit that is giving great joy to our partners and over 10,000 of their parents,"*

demonstrates tangible efforts to enhance well-being and support others.

In contrast, lay participants, particularly those with high emotional intelligence, demonstrated a tendency to overvalue affective cues, resulting in a high rate of false positives when compared to expert assessments. This pattern suggests that while high emotional intelligence enhances sensitivity to emotional language, it may simultaneously impair the ability to differentiate between superficial affective expressions and genuine markers of empathy.

Polarity:

This section investigates how participants identified text samples conveying positive vs. negative polarity. Similar to the previous two cases, Figure H3 displays the top word frequencies divided based on the agreement/ disagreement between expert and participant annotations:

1. Top-Left Plot (Expert=1, Participant=1): Correct Positive Annotations

In these samples, both participants and experts agreed the sentiment was positive. Notable words include profit, increas, net, rose, improv, growth, posit, and servic. These terms point to financial gains or upward trends (e.g., profit, increas, rose, improv, growth), reflecting favorable business outcomes. The presence of encouraging markers like posit and servic further highlights a generally optimistic message, which participants correctly identified.

2. Top-Right Plot (Expert=1, Participant=0): Missed Positive

Here, experts identified the text as positive, but participants labeled it as negative. Frequently highlighted words include loss, eur, increas, profit, sale, grew, oper, compar, agreement, rose, and sell. The presence of conflicting cues (e.g., loss versus grew or rose) might have confused participants. In some cases, the text may have referenced "losses narrowing" or "profits increasing," which would be net positive. Participants' attention to loss or other potentially negative terms (e.g., narrow, period) might have overshadowed positive signals like grew or rose.

3. Bottom-Left Plot (Expert=0, Participant=1): Incorrectly Labeled as Positive

Participants labeled these texts as positive, but experts judged them to be negative. Frequently highlighted words include eur, decreas, profit, oper, year, sale, compar, first, and compani. While some words (e.g., profit, increas) often signal positive sentiment, the overall context may have been negative (e.g., references to decreas, or disappointing performance compared to prior periods). Participants may have been swayed by individual upbeat terms, such as profit, without recognizing indications of downturn (decreas, losses, or other negative details implicit in the broader text).

4. Bottom-Right Plot (Expert=0, Participant=0): Correct Negative Annotations

Participants and experts both recognized these texts as negative. The top words—profit, year, net, loss, fell, eur, decreas, sale, declin, oper, lay, lower, drop, and quarter—strongly suggest downturns (e.g., loss, fell, decreas, declin, drop). Although words like profit and net appear, they likely reflect a smaller or declining profit context. This cluster clearly aligns with negative sentiment overall.

These patterns illustrate how business-related terms can be interpreted in ways that either inflate or obscure sentiment cues. Positive markers—like profit, growth, and rose—can lead participants to assume optimism even when the surrounding context (e.g., decreas, loss) suggests otherwise. Conversely, the mention of loss can overshadow signs of improvement (grew, narrowed losses). This analysis underscores the importance of considering context rather than single words when judging sentiment.

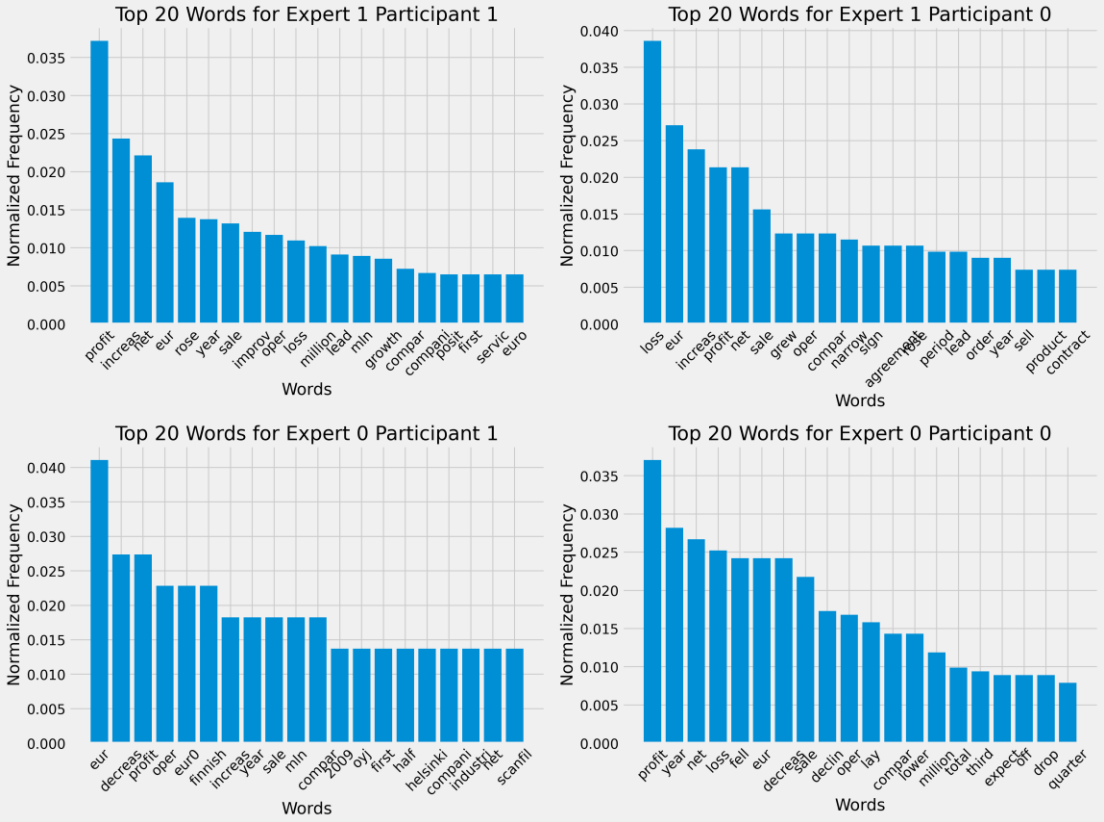

Figure H3. Normalized Word Frequencies for Each Group for Polarity Annotation

Appendix I: Using Cognitive – Affective Spectrum to Refine Training Data for FLMs

To explore the impact of annotator traits—specifically cognitive abilities and emotional intelligence—on the effectiveness of FLMs traditionally used in social science research, we conducted an additional analysis. This analysis employed annotations from both experts and non-experts as training data for finetuning pre-trained models. As reported in the manuscript, we had 240 samples, each annotated by 10 participants. After removing problematic responses (as outlined in the manuscript) we ended up with an average of 9.5 annotations per sample. For each sample, we selected annotations by participants with higher than median cognitive abilities for resilience, higher cognitive abilities and emotional intelligence for empathy, and higher emotional intelligence for polarity. In cases where a sample had more than one above-median annotation, we randomly selected one annotation. This resulted in 240 annotations for resilience and polarity and 234 annotations for empathy (some samples did not have any annotator with above median cognitive abilities and emotional intelligence). We then used this data to finetune the best pre-trained model (based on Table E) for each task. To ensure the best performance, we conducted hyperparameter optimization (the same process we reported in Appendix E) for each model. The final results are reported in Table 7 in the manuscript. It is worth noting that those results indicate that the overall performance of FLMs, even with expert annotations, was below expectations due primarily to reduced training sample sizes in this analysis (i.e., 240 vs. 1,482 training samples in resilience, 800 training samples in empathy, and 1,500 training samples in polarity).[5] However, our findings highlight that:

- FLMs finetuned with non-expert annotations from individuals with above-median cognitive abilities achieved performance levels comparable to those tuned with expert data for resilience tasks.

[5] We used the original training samples for hyperparameter optimization and testing.

- For empathy tasks, FLMs finetuned with annotations from non-experts with both above-median cognitive abilities and emotional intelligence showed slightly lower but still comparable performance to those tuned with expert data.
- FLMs finetuned with non-expert annotations from individuals with above-median emotional intelligence matched the performance of those tuned with expert data for polarity tasks.

Our results (in Table 7 of the main manuscript), aligned with our theorizing, advocate for assessing annotators' cognitive and affective capabilities and selecting those whose characteristics best match the cognitive-affective demands of the annotation tasks.

Appendix J: Analysis of Attention Weights

To provide further evidence that LLMs respond to enhancements in "cognitive abilities" and "emotional intelligence" while generating annotations, we conducted an attention-weight analysis. In this analysis, we track the amount of attention LLMs pay to phrases "cognitive abilities" and "emotional intelligence" as they generate the outputs (i.e., probability scores). Analyzing attention weights in Transformer models provides valuable insights into how these models process and represent information. Analyzing attention weights from the last layer of transformers offers significant advantages in understanding the semantic representations that models generate. This is because the last layer predominantly focuses on the most abstract features of the input data, capturing semantic meaning rather than mere linguistic syntax. The positional semantics embedded in the last layer's attention weights highlight key relationships and contextual cues that are essential for tasks such as sentiment analysis. Research indicates that the self-attention mechanism in the last layer of transformers encodes distinct semantic information, thereby allowing for richer interpretation of the data processed (Wu et al. 2020). In this context, attention scores from the last layer have been shown to correlate more directly with human interpretations of meaning, as demonstrated in tasks that require nuanced understanding of context, such as sentiment analysis and emotional inference. In contrast, earlier layers of transformer models typically encode syntactic features, which, while informative, do not capture the same depth of semantic information. It has been observed that attention patterns in earlier layers can emphasize syntactic relationships, but may lack the ability to represent the deeper contextual embeddings that are crucial for advanced language tasks (Mohebbi et al. 2023). As a result, focusing analysis on the last layer's attention weights allows researchers to gain insights into how models derive meaning from the input text, proving to be particularly beneficial for applications like multi-document summarization and contextual understanding (Hickmann et al. 2022).

If we can show that LLMs pay different levels of attention to "cognitive abilities" and "emotional intelligence" as they annotate resilience, empathy, and polarity samples, we can provide more evidence of the impact of this prompting strategy. To this end, we decided to conduct a detailed analysis of attention weights while the model generated the scores. However, since GPT4o is a closed LLM, we cannot gauge the attention weights from this model. Hence, we decided to use Meta's "Llama 3.1 70B Instruct," a popular open-source LLM with superb performance. This version of Llama has 80 attention layers with 64 attention heads within each attention layer. Given that the last attention layer is more related to semantics, we obtained attention weights from this layer for each token in "cognitive abilities" and "emotional intelligence" as the model generated each token of the output (e.g., "0.42"). Figure J provides an illustrative example regarding how we measured the attention weights. The plot displays the attention weights for "cognitive abilities" and "emotional intelligence" while the model generated each token in the output "\n0.42" as the probability score for a sentence to be high empathy. The first token in the output is "\n", the second one is "0", the third token is ".", and the fourth one is "42". For example, this model paid 0.03 units attention to phrase "emotional intelligence" and almost half of that to "cognitive abilities" while it generated token the first token ("\n"). We observed that the first and third tokens, which are punctuations in our specific analysis, had higher mean attention weights. While attention weights are not directly influenced by future tokens due to the causal nature of the attention mechanism, the model's representations and hidden states, which are used by the attention mechanism, can indirectly encode expectations about future tokens based on learned patterns in the training data. During the generation of a token (e.g., first token), the LLM's learned representations and hidden states, shaped by training data, reflect anticipations of likely continuations (second token). The self-attention mechanism then calculates attention

weights based on the current state and the preceding context, which indirectly incorporates these anticipations. This allows the model to generate token 0 in a way that is statistically consistent with observed sequential dependencies in the training data. This is consistent with the broader understanding of how LLMs learn and generate text by capturing statistical dependencies in language (Zhang et al. 2024). Therefore, instead of solely relying on the attention weights for the fourth token (e.g., 42), we considered using attention weights for both third and fourth tokens.

A key challenge in analyzing attention weights from LLMs is that they change depending on the specific words in each sentence. This means we can't directly compare attention weights for the same concept (like "cognitive abilities") across different sentences; it's not a fair comparison. To solve this, we used a consistent starting point. Across all our samples, the LLM always generated a newline character ("\n") followed by the digit "0" as the first two tokens. We used the attention weights assigned to these first two tokens as a baseline. Instead of comparing raw attention values for "cognitive abilities" and "emotional intelligence" across sentences, we measured how much the attention changed when the model generated the third and fourth tokens (which contains the actual probability score) relative to the attention it assigned to "\n" and "0". This approach lets us focus on how the model's attention to "cognitive abilities" and "emotional intelligence" increases or decreases as it generates the probability score for each sentence, effectively controlling for the influence of different sentence contexts, resulting in a much more accurate way to compare how the model attends to these concepts across our samples.

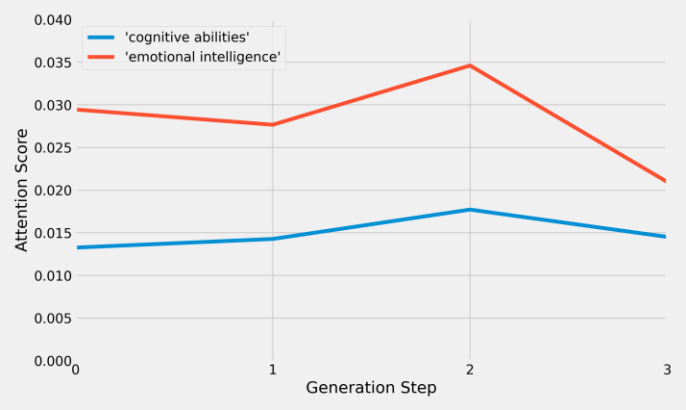

Figure J. LLM's Attention to "Cognitive Abilities" and "Emotional Intelligence"

After computing the attention weights during annotation of resilience, empathy, and polarity samples, we ran a logistic regression with LLM-Expert Match (equals 1 if LLM's annotation matched ground truth and 0 otherwise) as the DV and attention paid to cognitive abilities and emotional intelligence during the annotation as the independent variables. In addition, given that we are interested in associating attention weights for cognitive abilities and emotional intelligence to the type of task (e.g., annotation task that requires more cognitive processes than affective processes) as opposed to the difficulty of the task, we used Python package PyHard to compute and control for the annotation difficulty for each sample in our data. PyHard quantifies instance hardness by evaluating the local neighborhood structure of data points within the feature space, estimating the difficulty of classification based on factors such as class overlap, data sparsity, and the proximity of instances to decision boundaries.[6]

Appendix K: Systematic Review of Psychometric NLP Literature in Social Science Disciplines

To provide a comprehensive understanding of the application of traditional Natural Language Processing (NLP) techniques within the social sciences, a systematic literature review was conducted. This appendix details the rigorous methodology employed in this review, designed to identify and characterize the use of these methods in key social science disciplines. The objective was to contextualize our primary research by mapping

[6] Please see https://ita-ml.gitlab.io/pyhard/ for more details.

the landscape of NLP adoption in related fields, thereby highlighting the significance and positioning of psychometric NLP research.

The initial step involved identifying a representative set of social science disciplines. A multi-source approach was adopted to ensure a comprehensive and robust selection. We consulted a diverse array of authoritative sources, including:

- Major academic journal indexing services: Clarivate Journal Citation Reports, Google Scholar's journal categories, and Scimago Journal & Country Rank.
- Prominent social science organizations and repositories: the Social Science Research Network (SSRN), the Academy of Social Sciences (UK), and UK Research and Innovation (ESRC).
- Encyclopedic and community-curated resources: the Britannica and Wikipedia articles for "social science."

Table K1. Representative social science disciplines and counts of WoS articles from each citing seminal articles for lexicon based, CBM, and FLM methods

	Lexicons: LIWC (Tausczik and Pennebaker 2010)	**CBMs: LSTM** (Hochreiter and Schmidhuber 1997)	**FLMs: BERT** (Devlin et al. 2019)
Anthropology	1	1	1
Archaeology	0	2	4
Business	477	1220	750
Communication	186	5	49
Criminology & Penology	16	2	2
Economics	20	292	42
Education	83	52	124
Information Science & Library Science	104	244	529
Law	28	14	36
Linguistics	147	159	634
Political Science	54	5	33
Psychology	596	163	136
Sociology	31	9	15

Our synthesis identified thirteen representative social science disciplines for this review (see Table K1).

Subsequently, we selected a representative, highly-cited seminal paper for each of the three traditional NLP techniques examined in this manuscript: lexicon-based methods, custom-built models (CBMs), and fine-tuned masked language models (FLMs). The chosen studies were the most cited within their respective categories as detailed in our main manuscript.

- For **lexicon-based methods**, we selected Tausczik and Pennebaker (2010), the seminal work on LIWC, which had accrued 3,486 citations in the Web of Science (WoS).
- For **custom-built models (CBMs)**, we selected the foundational paper on LSTM networks by Hochreiter and Schmidhuber (1997), with 48,737 citations. While the Transformer architecture (Vaswani et al., 2017) is also a cornerstone of CBMs, its seminal paper was often co-cited with the paper for BERT. To avoid confounding citations related to CBMs with those for FLMs, we deliberately chose the second-most cited foundational CBM paper.
- For **fine-tuned masked language models (FLMs)**, we selected the paper introducing BERT by Devlin et al. (2019), which had 39,959 citations.

It should be noted that LLM (as the fourth NLP paradigm) was not included in this review. This is a deliberate exclusion, as their application in psychometric NLP is an emerging area with a limited number of studies to date, many of which are already discussed within the body of this dissertation.

Using Clarivate Web of Science (WoS), we identified all academic articles within the thirteen selected social science disciplines that cited these three seminal papers. This search yielded a corpus of 5,152 unique articles, detailed for each discipline in Table K1. There is significant variation among disciplines (part of which may be owed to individual disciplines' focus on academic outlets with more or less representation on Web of Science), but across most we identify a large representation of studies referencing these seminal works.

To gain a deeper insight into how these NLP methods are being utilized, we conducted a high-level characterization of the top five most-cited articles for each of the three seminal NLP papers within each of the thirteen disciplines. This resulted in a sample of 132 unique articles for detailed review, as some highly-cited papers were categorized under multiple WoS disciplines. In such cases of overlap, we proceeded to the next most-cited article to ensure a review of at least five distinct studies per discipline where possible.

The findings of this systematic review, as summarized in Table K2, demonstrate that a substantial portion of the most influential social science research leveraging these NLP techniques is dedicated to psychometric applications. Specifically, of the 132 articles analyzed, 106 employed NLP methods. Among these, 58 studies (44% of the total reviewed, or 55% of those using NLP) were identified as psychometric NLP. This significant representation underscores the importance and impact of applying NLP to measure psychological constructs within the broader social sciences, reinforcing the contributions of the present work.

Table K2. Summary of reviewed articles per social science discipline

Discipline	Total reviewed	NLP	Psych NLP	Study Type			
				Applied (A)	Conceptual (C)	Methods (M)	Survey (S)
Anthropology (Anth)	3	1	0	3	0	0	0
Archaeology (Arch)	4	3	2	4	0	0	0
Business (Bus)	19	10	7	11	3	6	2
Communication (Comm)	13	12	10	8	1	4	1
Criminology & Penology (Crim)	9	9	6	9	0	0	0
Economics (Econ)	17	9	6	10	4	3	2
Education (Educ)	14	8	6	12	2	1	1
Law (Law)	14	12	3	10	1	3	1
Information Sci. & Library Sci. (Libr)	13	9	4	7	1	6	2
Linguistics (Ling)	15	13	7	4	1	8	5
Political Science (Polit)	10	10	6	8	0	3	0
Psychology (Psych)	20	15	13	14	1	3	2
Sociology (Soc)	19	11	1	11	3	5	1
All reviewed (unique papers)	**132**	**106**	**58**	**81**	**18**	**33**	**8**

Table K3. Details of reviewed articles
Note 1: Y= yes/core component, P = partial/secondary component, N = No,
Note 1: Discipline and study type abbreviations defined in Table K2)

Citation	Title	Summary of study / NLP use	Discipline(s)	NLP?	Psych NLP?	Study Type
Abdi et al. 2019	Deep learning-based sentiment classification of evaluative text based on multi-feature fusion	Blends word-embedding, sentiment-lexicon and rule features in an LSTM framework to improve review-level sentiment prediction	Libr	Y	N	M
Abood & Feltenberger 2018	Automated patent landscaping	Semi-supervised text-analytics pipeline expands seed patents and prunes with ML to build topical patent sets	Law	Y	N	M, A
Alguliyev et al. 2019	The improved LSTM and CNN models for DDoS attacks prediction in social media	Uses Twitter sentiment streams and hybrid deep nets to forecast next-day DDoS activity	Polit	Y	P	M, A
Altmann & Mirkovic 2009	Incrementality and prediction in human sentence processing	Computational and behavioral work shows SRN models anticipate thematic roles during real-time comprehension	Psych, Ling	P	Y	A
Alva-Manchego et al. 2020	Data-Driven Sentence Simplification: Survey and Benchmark	Develops and evaluates automatic sentence-simplification methods	Ling	Y	N	M
Arseniev-Koehler & Foster 2022	Machine Learning as a Model for Cultural Learning: Teaching an Algorithm What it Means to be Fat	Uses word-embedding trajectories in news to reveal evolving weight stigma	Soc	Y	Y	A
Aziani et al. 2025	Conspiracy to Commit: Information Pollution, AI, and Real-World Hate Crime	1-D CNN links Google-search volumes for conspiracy terms to subsequent hate-crime counts	Crim	Y	N	A
Baden et al. 2022	Three Gaps in Computational Text Analysis Methods for Social Sciences	Maps methodological blind spots and future tasks for social-science NLP	Comm	Y	Y	C, M

Bae & Lee 2012	Sentiment Analysis of Twitter Audiences	Correlates tweet tone with influencer reach to gauge positive/negative sway	Libr	Y	Y	A
Barbado et al. 2019	A Framework for Fake Review Detection in Online Consumer-Electronics Retailers	Combines linguistic and behavioral cues to flag fraudulent reviews	Libr	Y	N	A
Bauer et al. 2023	Using NLP to Support Peer-Feedback in the Age of AI	Conceptual framework outlining NLP's role in generating and scaffolding student peer feedback	Educ	Y	P	C
Bazarova et al. 2013	Managing Impressions and Relationships on Facebook	LIWC dissects style and emotion in relational wall posts	Comm, Ling, Psych	Y	Y	A
Bi et al. 2020	Daily tourism volume forecasting for tourist attractions	LSTM integrates search-engine and weather series to predict daily visitors	Soc	N	N	M, A
Bi et al. 2021	Tourism demand forecasting with time series imaging	Converts arrival series to images, feeding CNN-LSTM for improved forecasts	Soc	N	N	M
Black et al. 2011	Emotions, Oral Arguments, and Supreme Court Decision Making	Links affective language in Supreme Court exchanges to case outcomes	Law	Y	Y	A
Block et al. 2019	A Personality Perspective on Business Angel Syndication	Infers investor Big-Five traits from tweets to study syndication patterns	Econ, Bus	Y	Y	A
Bojic 2024	Exploring the Socio-Technical Imaginary of AGI in Bard LLM	Narrative analysis of Bard interview uncovers optimistic vs. pessimistic AGI themes	Anth	N	N	A
Borovkova & Tsiamas 2019	An ensemble of LSTM neural networks for high-frequency stock market classification	Online-weighted LSTM ensemble classifies intraday price moves	Econ, Bus	N	N	M, A
Boyd & Schwartz 2021	Natural Language Analysis and the Psychology of Verbal Behavior	Survey of computational psycholinguistics and future NLP directions	Comm, Ling, Psych	Y	Y	S

Brueckner & Schuller 2015	Be at Odds? Deep and Hierarchical Networks for Conflict in Speech	Deep CNN-RNN system classifies conflict intensity from acoustic-linguistic cues	Comm	Y	Y	M
Burt & Reagans 2022	Team Talk: Learning, Jargon, and the Pulse of the Network	Tracks jargon drift and LIWC dimensions in team emails over time	Anth, Soc	Y	P	A
Calvo et al. 2017	NLP in Mental-Health Applications Using Non-Clinical Texts	Comprehensive review of sentiment and psych-signal extraction for mental health	Ling	Y	Y	S
Cao et al. 2023	How to Talk When a Machine Is Listening	Firms strategically down-weight negative tone in disclosures to fool sentiment algorithms	Econ, Bus	Y	Y	A
Caravale et al. 2023	Developing a Digital Archaeology Classification System	Topic modeling and NER build an ontology for digital-archaeology research	Arch	Y	N	A
Cavalieri et al. 2023	Intelligent Categorization of Italian Question-Time Documents	Benchmarks classic vs. deep text classifiers for agenda-topic labeling of parliamentary questions	Polit	Y	N	M
Chang & Masterson 2020	Using Word Order in Political Text Classification with LSTM	Shows LSTMs exploiting sequence improve policy-topic detection over bag-of-words	Polit	Y	N	M
Chen et al. 2017	Improving Sentiment Analysis via Sentence Type Classification	BiLSTM-CRF splits sentences by target count, then CNN predicts sentiment per type	Bus	Y	N	M
Chen & Cui 2020	Utilizing Student Time-Series Behaviour in LMS for Early Prediction	LSTM on click-stream sequences flags at-risk students weeks in advance	Educ	N	N	A
Chen et al. 2024	Using NLP to Evaluate Local Conservation Text	Topic and psychometric extraction from 624 UNESCO site reports identify integrity/authenticity themes	Arch	Y	Y	A
Choi et al. 2017	Using RNN Models for Early Detection of Heart-Failure Onset	GRU network mines temporal EHR events for 12–18 month incident HF prediction	Libr	N	N	A

Chung & Pennebaker 2011	Using Computerized Text Analysis to Assess Threats	LIWC reveals linguistic markers distinguishing serious from hoax threats	Comm, Crim	Y	Y	A
Collins et al. 2015	Influence of Amicus Briefs on Supreme Court Opinion Content	Text-reuse analysis shows which brief traits drive doctrinal uptake	Law, Soc	Y	N	A
Condevaux & Mussard 2024	Automated NLP: Transformers and Challenges for Economics	Explains attention, LLM training phases, and future transformer uses for economists	Econ	N	N	C
Corley & Wedeking 2014	The (Dis)Advantage of Certainty: Legal Language	Higher certainty wording in opinions increases lower-court reliance	Law, Soc	Y	N	A
Crabtree et al. 2020	It Is Not Only What You Say… Campaign Sentiment	Analyzes emotional tone in manifestos to predict strategic positioning	Polit	Y	Y	A
Davidson 2023	Start Generating: Harnessing Generative AI for Sociology	Commentary on LLM opportunities and caveats for sociological inquiry	Soc	N	N	C
Degli Esposti et al. 2020	Use of Copyrighted Works by AI Systems: Art Works in the Data Mill	Legal analysis of training-data reuse and similarity metrics for generative AI	Law	N	N	C
Drouin et al. 2017	Linguistic Analysis of Chat Transcripts from Predator Stings	Psycholinguistic differences between offenders and undercover agents	Crim	Y	Y	A
Ellefsen et al. 2019	Semi-Supervised Deep Architecture for Turbofan RUL	Combines unsupervised pre-training and GA hyper-search for life-prediction accuracy	Bus	N	N	M, A
Evans & Aceves 2016	Machine Translation: Mining Text for Social Theory	Surveys text-mining workflows to build sociological theory	Soc	Y	P	S
Fan et al. 2023	Multimodal Knowledge Graph of Chinese Operas	Builds ontology, multimodal KG and multi-task model for sentiment/genre recognition	Arch	Y	Y	A
Fernandes et al. 2020	Appellate Court Modifications Extraction for Portuguese	BiLSTM-CRF tags modification provisions in Brazilian appeals with 95 % F1	Law	Y	N	M

Fotiadis et al. 2021	The Good, the Bad and the Ugly on COVID-19 Tourism Recovery	LSTM and GAM scenarios project 30–76 % drop in arrivals post-pandemic	Soc	N	N	A
Fujimoto et al. 2022	Integrated Molecular & Affiliation Network Analysis	Combines gene and social networks; cites LLM transfer learning conceptually	Anth, Soc	N	N	A
Gillani et al. 2023	Unpacking the Black Box of AI in Education	High-level overview of LLM use, risks, and governance in educational contexts	Educ	N	N	C
Guo & Wang 2024	ChatGPT's Potential to Support Teacher Feedback in EFL Writing	Small-scale experiment assesses LLM-generated feedback quality and teacher uptake	Educ	N	N	A
Guo et al. 2022	A Survey on Automated Fact-Checking	Comprehensive review of datasets, architectures, and open challenges	Ling	Y	N	S
Halford et al. 2022	Anti-Social Behaviour in the Pandemic	NLP classifies police text records to map lockdown-era ASB trends	Crim	Y	N	A
Han et al. 2024	Hygrothermal Performance of Enclosures and Energy Efficiency	Hybrid physics-plus-AI model predicts storage-room microclimate to relax RH controls	Arch	N	N	A
Hanauer et al. 2012	Project Ownership in Undergraduate Research	LIWC shows language of ownership grows with research experience	Educ	Y	Y	A
Harackiewicz et al. 2016	Closing Achievement Gaps with Utility-Value Essays	Essay-language analysis links cognitive words to performance gains across races	Psych	Y	Y	A
Hewamalage et al. 2021	Recurrent Neural Networks for Time-Series Forecasting	Large empirical study and best-practice guide for RNN forecasting	Econ, Bus	N	N	S
Hickman et al. 2022	Automated Video-Interview Personality Assessments	Multimodal transformer predicts HEXACO traits; validation against human ratings	Psych	Y	Y	M, A
Ho et al. 2018	Psychological, Relational, and Emotional Effects of	Content and survey data show emotional benefits after sharing with a bot	Comm	Y	Y	A

	Chatbot Self-Disclosure					
Hoover et al. 2020	Moral Foundations Twitter Corpus	Releases 35 k tweets hand-labeled for ten moral sentiments to aid NLP	Psych	Y	Y	M
Hu et al. 2020	Graph Neural News Recommendation	Heterogeneous GNN models long- vs. short-term interests for personalized news	Libr	Y	N	M
Hu et al. 2021	Transaction-Based Classification of Ethereum Smart Contracts	LSTM on sliced transaction sequences detects malicious and duplicate contracts	Libr	N	N	A
Iliev et al. 2015	Automated Text Analysis in Psychology	Reviews methods, apps, and future of computational psycholinguistics	Ling, Psych	Y	Y	S
Ireland et al. 2011	Language Style Matching Predicts Relationship Initiation and Stability	Word-level synchrony foresees dating success and long-term stability	Psych	Y	N	A
Jeong et al. 2020	Context-Aware Citation Recommendation with BERT & GCN	Hybrid graph+transformer suggests papers from manuscript context sentences	Libr	Y	N	M, A
Jiang et al. 2020	How Can We Know What Language Models Know?	Probing shows factual gaps and surface-cue reliance in pre-trained LMs	Ling	Y	N	M
Jiang et al. 2021	Polarization Over Vaccination	Topic-sentiment analysis of vaccine tweets reveals ideology-specific hesitancy frames	Comm	Y	Y	A
Joksimovic et al. 2015	Social Presence in MOOCs as Predictor of Performance	Content-analysis metrics of presence correlate with course grades	Educ	Y	Y	A
Joksimovic et al. 2018	How Do We Model Learning at Scale?	Systematic review of MOOC analytics, highlighting NLP gaps	Educ	P	N	S
Jones 2016	Talk Like a Man: Hillary Clinton, 1992-2013	Tracks shifts toward masculine and emotional language over two decades	Polit	Y	Y	A
Kaminski & Hopp 2020	Predicting Crowdfunding Outcomes with Multimodal Signals	Sentiment, uncertainty, and imagery cues forecast Kickstarter success	Econ, Bus	Y	Y	A

Kim et al. 2022	Design Principles and Architecture of an L2-Learning Chatbot	Presents multimode voice chatbot "Ellie" and pilot usability study	Educ, Ling	Y	N	M, A
Kolbel et al. 2024	Ask BERT: Climate-Risk Disclosure and CDS Term Structure	BERT measures transition/physical climate language to explain CDS spreads	Econ, Bus	Y	N	A
Kozlowski et al. 2019	Geometry of Culture: Meanings of Class via Word Embeddings	Diachronic embeddings map changing class connotations in literature	Soc	Y	N	A
Krueger & Dayan 2009	Flexible Shaping: How Learning in Small Steps Helps	Neural-network simulations show staged shaping speeds sequence learning	Psych	N	N	A
Lavorgna et al. 2020	FloraGuard: Tackling Illegal Online Plant Trade	Keyword and topic models detect endangered-flora sales on forums	Crim	Y	N	A
Lazer & Radford 2017	Data ex Machina: Introduction to Big Data	Commentary on big-data promises and perils for social research	Soc	P	N	C
Le Mens et al. 2023	Using ML to Uncover Semantics of Concepts	BERT genre-typicality scores correlate > 0.85 with human ratings	Soc	Y	N	A
Li et al. 2020	Attention-BLSTM for Multimodal Emotion Recognition	Learns temporal EEG/physio patterns; decision-level fusion predicts valence/arousal	Libr	N	Y	M
Li et al. 2021	Measuring Corporate Culture Using Machine Learning	Topic model and sentiment extraction from earnings calls build culture indices	Econ, Bus	Y	Y	A
Liu et al. 2022	Automated Detection of Engagement in MOOC Discussions	Transformer detects emotional/cognitive states and predicts achievement	Educ	Y	Y	A
Loos et al. 2023	Using ChatGPT in Education: Human Reflection on ChatGPT's Self-Reflection	SWOT dialogue with GPT-3 highlights hallucination and teaching implications	Soc	N	N	M

Lund et al. 2023	ChatGPT and a New Academic Reality	Ethical analysis of LLM-generated manuscripts and peer-review challenges	Libr	N	N	C
Major et al. 2014	The Ironic Effects of Weight Stigma	Speech analyses show coping mechanisms backfire under stigma priming	Psych	Y	Y	A
Mandal et al. 2021	Unsupervised Similarity Between Legal Case Reports	Embedding-based clustering retrieves precedent cases without labels	Law	Y	N	A
Markowitz & Griffin 2020	When Context Matters: Styles of Truth vs. Lies	LIWC shows deception style varies by genre, not just truthfulness	Crim, Law, Psych	Y	Y	A
Masini et al. 2023	Machine Learning Advances for Time-Series Forecasting	Survey of penalized, ensemble, deep, and tree models, with economic examples	Econ, Bus	N	N	S
Matsumoto & Hwang 2015	Word Usage by Truth Tellers vs. Liars	Lexical patterns in mock-crime interviews differentiate deception	Crim, Psych	Y	Y	A
Morina & Ybarguen 2024	Statistical Modelling for COVID-19 Incidence in Spain	Bayesian ARCH model reconstructs under-reported cases and forecasts scenarios	Comm	N	N	A
Mubarak et al. 2021	Predictive Learning Analytics Using Deep Learning in MOOC Videos	LSTM on video click-streams forecasts weekly performance (82–93 % accuracy)	Educ	N	N	A
Nelson 2020	Computational Grounded Theory: A Framework	Merges unsupervised NLP with iterative human coding to build theory	Soc	Y	P	C
Nelson et al. 2021	Future of Coding: Hand-Coding vs. Computer-Assisted Methods	Benchmarks dictionary, topic, embedding, and BERT against manual labels	Soc	Y	N	A
Nguyen et al. 2018	RNN-Based Models for Recognizing Requisite & Effectuation Parts	BiLSTM-CRF plus cascading models tag overlapping legal provisions	Law	Y	M	M
Obschonka & Fisch 2018	Entrepreneurial Personalities in Political Leadership	Twitter-based personality assessment compares CEOs and politicians	Econ, Bus	Y	Y	A

Olive et al. 2019	One-Size-Fits-All Neural Network for At-Risk Students	Peer-context features improve neural early-warning across 5,487 courses	Educ	N	N	A
Onan 2020	Mining Opinions from Instructor Evaluation Reviews	Attention-RNN with GloVe achieves 98 % accuracy on 154 k SET reviews	Educ	Y	Y	A
Owens & Wedeking 2011	Justices and Legal Clarity	Automated complexity scores of opinions predict reversal and citation	Law, Soc	Y	N	A
Park et al. 2015	Automatic Personality Assessment Through Social Media Language	Open-vocabulary model predicts Big-Five with > .3 correlations	Psych	Y	Y	A
Pater 2019	Generative Linguistics and Neural Networks at 60	Historical essay urging fusion of deep learning and generative theory	Ling	P	N	C
Pitt et al. 2002	Toward a Method of Selecting Among Computational Models	Introduces MDL as principled metric for cognitive-model comparison	Psych	N	N	C
Prufer & Prufer 2020	Data Science for Entrepreneurship Research	NLP of Dutch job ads tracks changing skill demand	Econ, Bus	Y	Y	C, A
Rauthmann et al. 2014	The Situational Eight DIAMONDS	Defines taxonomy; demonstrates text-based coding of situation features	Psych	P	Y	M
Rees et al. 2013	Narrative, Emotion and Action in Memorable Dilemmas	Emotion-word analysis explains which dilemmas students recall most	Educ	Y	Y	A
Relins et al. 2025	Using Instruction-Tuned LLMs to Identify Vulnerability in Police Narratives	GPT-style models replicate human coding of mental health and homelessness	Crim	Y	Y	A
Renz et al. 2018	Two Strategies for Qualitative Content Analysis	Shows intramethod triangulation combining NLP outputs with manual coding	Libr	Y	Y	M
Rheault et al. 2019	Incivility Toward Women Politicians on Social Media	Classifier detects gendered slurs and threats in tweets to MPs	Polit	Y	Y	A

Richardson et al. 2014	Language Style Matching and Police Interrogations	Style alignment predicts likelihood of confession in real cases	Law, Psych	Y	N	A
Rojas-Barahona 2016	Deep Learning for Sentiment Analysis	Overview of CNN/LSTM sentiment models and linguistic challenges	Ling	Y	N	S
Roy 2024	Detecting Sarcasm in Code-Mixed Social Posts	BERT-LSTM ensemble achieves 96 % accuracy on Hinglish sarcasm corpus	Soc	Y	N	M
Salinas et al. 2020	DeepAR: Probabilistic Forecasting with Autoregressive RNNs	Trains global RNN on many series to outperform classical demand models	Econ, Bus	N	N	M
Shapiro et al. 2022	Measuring News Sentiment	Creates domain-robust financial-news sentiment lexicon and index	Econ, Bus	Y	Y	A
Shin et al. 2020	Enhancing Social-Media Analysis with Visual Analytics	Integrates NLP, image-tags, and dashboards for crisis monitoring	Libr	Y	N	M, A
Song et al. 2019	Semantic Neural Machine Translation Using AMR	Injects AMR graphs into seq-to-seq MT improving BLEU on En-De	Ling	Y	N	M
ten Brinke & Porter 2012	High-Stakes Interpersonal Deception	Multimodal study links verbal cues and behavioural signals to guilt/innocence	Law, Psych	Y	Y	A
Toma & D'Angelo 2015	Linguistic Cues for Expertise in Online Medical Advice	Style and anxiety markers signal advisor credibility	Comm, Ling, Psych	Y	Y	A
Toma & Hancock 2012	Linguistic Traces of Deception in Dating Profiles	LIWC shows deceivers use fewer first-person and more negations	Comm	Y	Y	A
Tumasjan et al. 2011	Election Forecasts With Twitter	Party-name sentiment aggregates predict German election shares	Libr	Y	Y	A
Viehmann et al. 2023	Stance Classification of Opinions on Policy Measures	Benchmarks BERT vs. feature models on German COVID-19-policy tweets	Comm	Y	N	M
Villata et al. 2022	Thirty Years of Artificial Intelligence and Law: The Third Decade	Review notes ML shift toward text-based legal analytics	Law	N	N	S

Vuong et al. 2023	SM-BERT-CR: Deep Learning for Case-Law Retrieval	BERT with contrastive loss ranks precedents and supplies explanations	Law	Y	N	A
Wahman et al. 2021	From Thin to Thick Representation	Topic model of Malawi parliamentary debates shows women MPs' agenda expansion	Polit	Y	N	A
Walczak & Cellary 2023	Challenges for Higher Education in the Era of Generative AI	Outlines institutional, ethical, and pedagogical issues GenAI poses	Econ, Bus	N	N	C
Wang et al. 2020	What Influences Sales Market of New-Energy Vehicles?	Text mining of consumer comments extracts purchase motives	Econ, Bus	Y	N	A
Warstadt et al. 2020	BLiMP: Benchmark of Linguistic Minimal Pairs	67 k minimal-pair dataset evaluates grammar knowledge in LMs	Ling	Y	N	M
Webster 2018	Anger and Declining Trust in Government	Linguistic anger scores from open-ended survey answers predict trust erosion	Polit	Y	Y	A
Welbers et al. 2017	Text Analysis in R	Hands-on tutorial of topic, dictionary, and network methods	Comm	Y	P	M
Wiese et al. 2020	Quant GANs: Deep Generation of Financial Time Series	TCN-based GAN generates risk-neutral synthetic asset paths	Econ, Bus	N	N	M
Willemsen et al. 2011	Psychopathy and Lifetime Experiences of Depression	Interview language reveals affective patterns linked to psychopathy	Crim	Y	Y	A
Wojcieszak et al. 2023	No Polarization From Partisan News	Longitudinal trace data show limited attitude shifts despite partisan exposure	Comm, Polit	Y	Y	A
Wolff et al. 2016	Teacher Vision: Expert vs. Novice Perception	Think-aloud transcriptions analyzed for situation awareness cues	Educ, Psych	Y	Y	A
Xu et al. 2023	Forecasting Daily Tourism Demand with Multiple Factors	Temporal-fusion encoder with Bayesian tuning provides interpretable forecasts	Soc	N	N	M, A
Yan et al. 2022	Reinforcement Learning for Logistics and SCM	Survey of RL algorithms and supply-chain applications	Econ, Bus	N	N	C

Zemblys et al. 2019	gazeNet: End-to-End Eye-Movement Event Detection	CNN classifies fixations, saccades, PSOs directly from raw gaze streams	Psych	N	N	M
Zhang et al. 2018	LSTM for Machine Remaining Life Prediction	Maps sensor data to health index then forecasts turbofan RUL	Bus	N	N	M
Zhang & Ghorbani 2020	Overview of Online Fake News	Reviews characterization, detection and challenges of fake-news research	Libr	Y	N	S
Zhao & Li 2022	Fuzzy or Clear? Computational Modeling of L2 Lexical–Semantic Representation	DevLex-II simulations show AoA effects on fuzzy bilingual categories	Ling	Y	Y	M
Zhong et al. 2023	Robust Crisis Communication in Turbulent Times	Topic-emotion analysis of US governors' COVID tweets reveals strategy clusters	Polit	Y	Y	A

Please note that due to the long list of references in Table K3, we can provide the full list of papers upon request.

Appendix L: Systematic Review of Resilience, Empathy, & Polarity in Social Science Disciplines

1. Resilience in Social Science Research

We conducted a structured search in the Web of Science Core Collection to locate scholarship on "resilience." The term "resilience" was specified in the author-keyword field, ensuring that each retrieved record explicitly framed the construct as central. To guard against papers that deploy the same term in enterprise, organizational, network, or other technical settings, we applied an exclusion filter across titles, abstracts, and author keywords that removed records containing "system*," "network*," "supply chain*," "infrastructur*," "organization*/organisation*," "urban*," "regional*," "ecological," "cyber," and similar variants. These exclusions were informed by a preliminary scoping exercise that revealed persistent false positives in areas such as network resilience, supply-chain resilience, and urban or ecological resilience, all of which lie outside the scope of individual psychology.

We then restricted the results to the thirteen social-science disciplines listed in Appendix K—for example, anthropology, business, sociology, psychology, and communication—thereby maintaining disciplinary focus while retaining relevant interdisciplinary work. This two-step strategy balanced sensitivity and precision: it captured diverse conceptualizations of personal resilience yet eliminated the extensive systems-level literature that would otherwise blur our analytic lens. The final query yielded more than 4,027 distinct studies cited in 61,638 articles. Figure L1 presents the distribution of publications across Web of Science categories, with education, psychology, and business accounting for the largest shares. Figure L2 charts publication and citation counts from 1996 to 2025, revealing a sharp rise over the past decade and underscoring the growing prominence of resilience within social-science research.

Figure L1. Resilience Related Publications in Social Science by WoS Categories

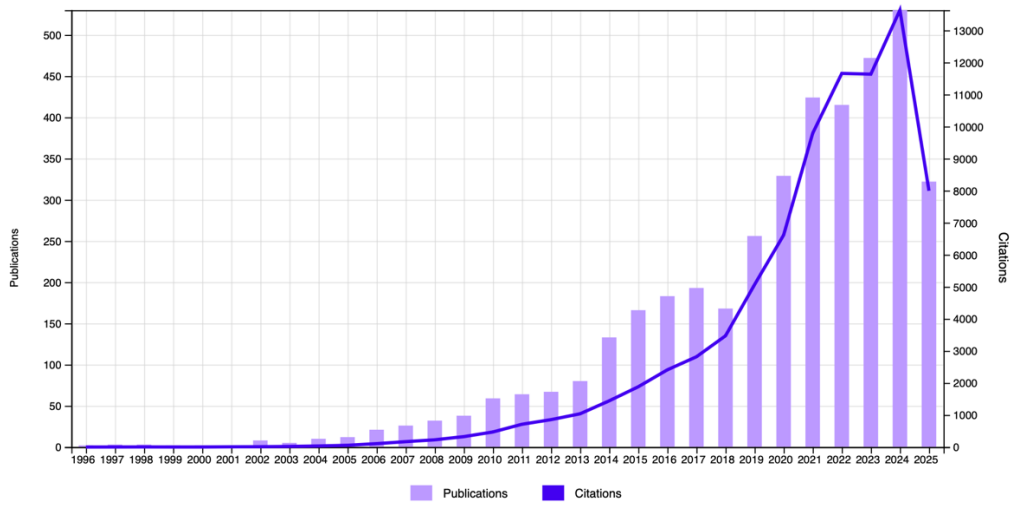

Figure L2. Resilience Related Publications & Citations in Social Science Research

2. Empathy in Social Science Research

We conducted a similar search in the Web of Science Core Collection to locate scholarship on "empathy." The term "empathy" was specified in the author-keyword field, ensuring that each retrieved record explicitly framed the construct as central. We then restricted the results to the thirteen social-science disciplines listed in Appendix K. This search resulted in 4,358 publications and 74,286 citations.

Figure L3 presents the distribution of publications across Web of Science categories, with education, psychology, and business accounting for the largest shares. Figure L4 charts publication and citation counts from 1990 to 2025, revealing a sharp rise over the past decade and underscoring the growing prominence of empathy within social-science research.

Figure L3. Empathy Related Publications in Social Science by WoS Categories

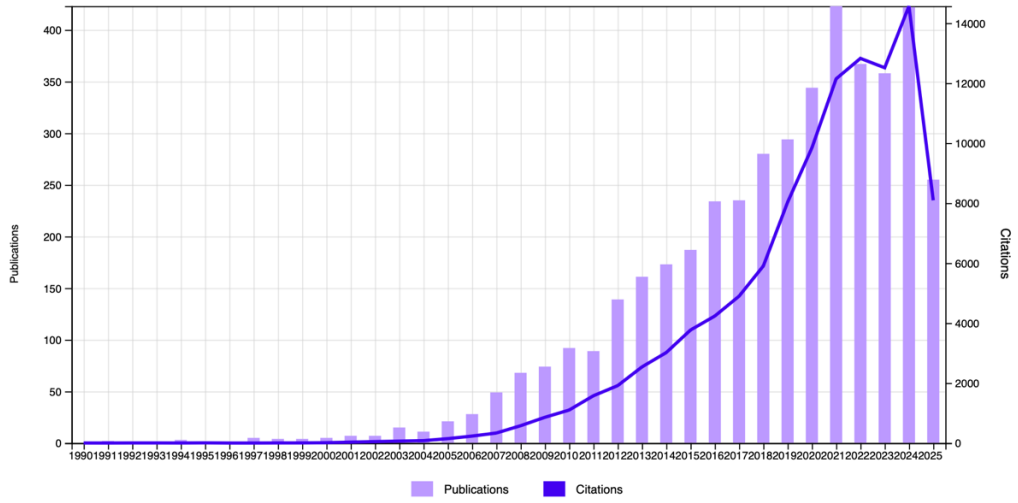

Figure L4. Empathy Related Publications & Citations in Social Science Research

3. Polarity in Social Science Research

We conducted a parallel search in the Web of Science Core Collection to map scholarship on evaluative polarity. Instead of querying the keyword "polarity," we specified "sentiment" in the author-keyword field. This substitution minimizes noise: polarity is routinely invoked in physics, chemistry, electrical engineering, and studies of political polarization, whereas sentiment is the preferred label for the positive–negative valence construct at the heart of computational linguistics and psychometrics. Using "sentiment" therefore targets the intended conceptual domain while avoiding large volumes of irrelevant literature. As in the resilience/ empathy search, we confined results to the thirteen social-science disciplines listed in Appendix K, preserving disciplinary focus without sacrificing coverage of interdisciplinary work.

The refined query yielded 4,516 publications that have been cited 52,827 times. Figure L5 displays the disciplinary distribution, with business, information science, and communication producing the greatest shares.

Figure L6 charts annual publication and citation counts from 1992 to 2025, revealing a pronounced upswing in the past decade and underscoring the growing prominence of polarity research within the social sciences.

Figure L5. Polarity Related Publications in Social Science by WoS Categories

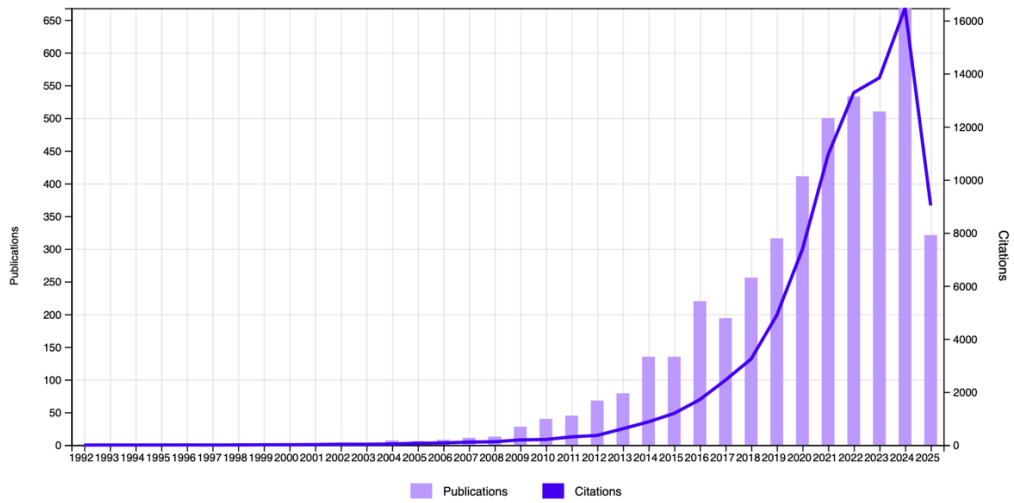

Figure L6. Polarity Related Publications & Citations in Social Science Research

Our analysis indicates that, over the past three decades, more than 12,000 social-science publications have treated resilience, empathy, and polarity as focal constructs. The annual volume of these studies —and their citation counts—has climbed steeply during the last ten years, underscoring the accelerating scholarly attention each construct now commands.

Appendix M: Theoretical Framework

1. Dual-Process Theory

Large corpora of text data labeled with psychological constructs has provided the lifeblood for psychometric NLP methods for decades. Human annotators with a variety of expertise and training are typically relied on to provide these seed labels through manual examination of these corpora. The process of human annotation is not a uniform or mechanical task. Rather, it is a complex act of human judgment, susceptible to a range of influences including cognitive load and the inherent characteristics of the annotator. Dual-Process Theory (DPT) characterizes human judgment as the interplay of two qualitatively different modes: System/Type 1 processes that are fast, automatic, associative, and high capacity, and System/Type 2 processes that are slower, controlled, rule-based, and capacity-limited (Evans 2008, Evans and Stanovich 2013). In the classical DPT view, a rapid Type 1 response is generated by default; when sufficient resources and motivation are available, a deliberative Type 2 process may intervene to evaluate, revise, or override that initial response. This framework organized influential programs of research on belief bias, heuristic judgment, and reasoning error (Evans and Stanovich 2013).

A key refinement in DPT created a more direct bridge to measurable individual differences by functionally defining Type 2 processes through their dependence on working memory and Type 1 processes as autonomous (De Neys 2018). This reframing solidifies the link between Type 2 processing and cognitive ability, a connection long established by the theory's leading proponents. For instance, Evans (2003) argued that since "System 2 requires working memory whose capacity is known to vary across individuals," its function "should be related to measures of general intelligence" (p. 456), a view later reinforced by the observation that Type 2 processes are "correlated with cognitive ability" (Evans & Stanovich, 2013, p. 225). This theoretical position is strongly substantiated by a robust, independent body of research in the working-memory literature, which consistently documents strong correlations (often $r>.50$) between working-memory capacity and fluid intelligence (Shipstead, Harrison, and Engle, 2016). At the same time, this framework connects autonomous Type 1 processes to emotional intelligence. Theorists propose that core aspects of emotional functioning, such as "some processes of emotional regulation," fall into the category of autonomous processes (Evans & Stanovich, 2013, p. 236). Fiori's (2009) pivotal dual-process analysis of emotional intelligence elaborates on this, proposing that individual differences in emotional intelligence reflect, in part, the efficiency of automatic, Type-1-like emotional processing. This link is supported by contemporary findings, such as evidence that emotional intelligence predicts early, low-effort attentional biases toward emotional stimuli, a plausible mechanism for shaping autonomous affective responses (Suslow et al., 2022).

Complementing DPT research on individual differences, Hot–Cool System theory of self-regulation (Metcalfe & Mischel, 1999; Mischel & Ayduk, 2004) shows that emotionally "hot" cues trigger fast, stimulus-bound responses, whereas "cool" processing supports symbolic re-representation and regulation. Research on hot vs. cool executive functions further documents that tasks high in affective salience preferentially recruit hot systems (analogous to Type 1 in DPT), while abstract, decontextualized problems rely on cool control resources (analogous to Type 2 in FPT).

Crucially, the distinction between these systems does not imply that they operate in a mutually exclusive manner. A central tenet of DPT is that an autonomous Type 1 process generates a default, intuitive response to a stimulus. Engaging in a task that requires careful thought or "paying attention" does not simply switch this initial process off. Instead, the deliberative Type 2 system may then intervene to evaluate, and potentially override, that initial impression, provided the individual has sufficient motivation and cognitive resources. As influential research has shown, even when individuals are explicitly instructed to be careful and deliberate, their final judgments are often anchored by immediate impressions and heuristics, especially under conditions of cognitive load or ambiguity (Kahneman, 2011; Stanovich & West, 1998). Thus, even within structured, rule-governed tasks, Type 1 processing provides the initial, often affective, "raw material" for judgment, which Type 2 processes then refine, correct, or endorse.

In sum, DPT provides a principled framework for linking distinct modes of information processing to specific individual differences. Type 2 processing, which is deliberative and analytical, is governed by the resources supplied by an individual's Cognitive Ability. In contrast, Type 1 processing, which is intuitive and automatic,

is guided by an individual's proficiency in interpreting and utilizing affective information, a key component of Emotional Intelligence.

2. Cognitive – affective Spectrum

Language processing is a multifaceted activity that engages a range of cognitive and affective processes (Pennebaker & Francis, 1996; Panksepp, 2003). On the cognitive side, attention, perception, and reasoning support analytical tasks, including problem-solving and interpretation of complex narratives (Panksepp, 2003). On the affective side, emotional resonance and sentiment sensitivity shape how individuals experience and respond to various cues (Davidson & Sutton, 1995). Various psychological constructs may be processed utilizing a combination of these aspects, each depending on different blends of cognitive and affective engagement. For this purposes of examining the impact of cognitive and affective processes, we have identified three constructs which relate to them in contrasting ways – resilience, which is highly cognitive in nature, polarity, which is highly affective, and empathy, which relates significantly to both.

2.1. Resilience Annotation:

The annotation of psychological resilience is fundamentally a task of complex cognitive appraisal, demanding deliberative, analytical thought that places it firmly at the cognitive end of our proposed spectrum. Psychological resilience is not a static trait but a dynamic process involving positive adaptation in the face of significant adversity (Masten 2001).

The theoretical and empirical literature consistently identifies cognitive appraisal as a core mechanism of the resilience process. Following the seminal work of Lazarus and Folkman (1984), this appraisal involves two stages: a primary appraisal that assesses the threat posed by a stressor, and a secondary appraisal that evaluates an individual's resources and ability to cope (Folkman et al. 1986). Consequently, an annotator tasked with judging resilience from a text must engage in a parallel evaluative process. This requires them to: (1) identify and comprehend the nature of the adversity or stressor described in the narrative; (2) analyze the subject's response to that adversity, searching for evidence of active coping strategies, problem-solving behaviors, and cognitive reappraisal (e.g., reframing the negative event, finding purpose, or fostering optimism); and (3) synthesize these disparate elements into a holistic judgment about the degree of resilience demonstrated (Riepenhausen et al. 2022).

This process of deconstruction, analysis, and synthesis is a form of top-down, effortful cognition (Dajani & Uddin 2015). It heavily taxes fluid cognitive abilities—the capacity to reason, problem-solve, and interact with novel information. The Cognitive Appraisal of Resilience (CAR) model provides a direct theoretical link, proposing that cognitive functions such as cognitive flexibility and executive functioning are critical moderators that mitigate the negative impact of adverse events (Yao & Hsieh 2019). A key feature of this cognitively demanding task is its reliance on what has been termed "hypothetical thinking" (Evans 2003) or "cognitive decoupling" (Evans and Stanovich 2013). The annotator must construct and maintain a mental model of the text author's situation, simulating their mental state and psychological trajectory to evaluate their response. This act of holding a complex simulation in mind, separate from one's own reality, is a hallmark of higher-order cognition and places a significant load on working memory. Therefore, resilience annotation is positioned at the cognitive pole of the spectrum, representing a deliberative, analytical task that depends on the cognitive systems responsible for abstract reasoning, problem-solving, and mental simulation.

2.2. Polarity Annotation:

In stark contrast to the deliberative nature of resilience annotation, the task of annotating text for polarity—that is, general sentiment—is dominated by rapid, intuitive, and affective judgment. This places it near the affective terminus of our proposed spectrum. Research in judgment and decision-making has extensively documented the role of the "affect heuristic," a mental shortcut wherein individuals use their immediate emotional reactions—their feelings of "goodness" or "badness"—as a proxy for more effortful evaluation (Slovic et al. 2007). Polarity annotation is a quintessential instantiation of this heuristic in action. The task requires an annotator to form a gestalt impression of the overall emotional valence of a text, a judgment that is typically not derived from a slow, logical decomposition of syntax and semantics.

Instead, the judgment of polarity relies on an immediate "gut feeling" or an emotional resonance triggered by the affective cues embedded within the text (Finucane et al. 2000). This process is characteristically fast,

automatic, and operates largely below the threshold of conscious, voluntary control, aligning perfectly with descriptions of associative and intuitive thought (Evans & Stanovich 2013). The "affect-as-information" model further clarifies this mechanism, positing that individuals often treat their feelings as a direct source of information to guide judgment (Clore & Huntsinger 2007). When annotating for polarity, the annotator implicitly asks, "How does this text make me feel?" The resulting affective state—positive, negative, or neutral—becomes the primary data point for the annotation. This entire operation is driven by what is often termed the "hot" or associate-processing system, which is quick, automatic, and emotional (Clore & Huntsinger 2007). Unlike resilience annotation, this task requires minimal cognitive decoupling; it is not about analyzing an external situation but about registering an internal, intuitive reaction.

Thus, polarity annotation occupies the affective end of the spectrum. It is a task that relies on the rapid, automatic, and intuitive assessment of emotional cues, making successful performance heavily dependent on an individual's sensitivity to and ability to accurately interpret affective information.

2.3. Empathy Annotation:

Positioned between the poles of resilience and polarity, the annotation of empathy represents a hybrid task, the successful execution of which requires a sophisticated blend of both cognitive and affective processing. There is a robust consensus in the psychological literature that empathy is a multidimensional construct comprising at least two distinct, yet interacting, components (Decety & Jackson, 2004; Singer & Lamm, 2009).

The first component is affective empathy, which involves the capacity to vicariously share or resonate with another person's emotional state, a phenomenon also known as "emotional contagion" (Hatfield et al. 1994). This process is more automatic, less cognitively demanding, and constitutes the ability to feel with another person (Singer & Lamm, 2009).

The second component is cognitive empathy, defined as the capacity to understand and adopt another person's psychological perspective—a process often referred to as "Theory of Mind" or mentalizing. This is a rational, controlled process that is closely associated with executive functioning and requires the deliberative effort to imagine and comprehend another's thoughts and feelings. It is the ability to know what another person is feeling (Frith & Frith, 2006; Decety & Jackson, 2004).

Accurately annotating a text for empathy necessitates the engagement of both of these systems (Singer & Lamm, 2009). An annotator must first deploy cognitive empathy to deconstruct the narrative, understand the context of the situation described, and form a coherent mental model of the subject's internal state. Subsequently, or in parallel, the annotator must engage affective empathy to gauge the degree of emotional resonance, concern, and warmth expressed in the text. A text might, for example, demonstrate a flawless cognitive understanding of another's predicament but betray a cold, detached emotional stance, which would warrant a low empathy rating. Conversely, a text might overflow with raw emotion but lack any clear understanding of the other's unique perspective, also indicating a deficit in overall empathy.

Dual-process models of empathy provide neurological and cognitive support for this distinction, showing that cognitive tasks like perspective-taking and affective tasks like emotional resonance are supported by distinct neural activations that work in concert (Singer & Lamm, 2009). A successful empathy annotator, therefore, cannot rely solely on logical deduction or on pure emotional reaction; they must be capable of integrating both. This task requires a state of partial cognitive decoupling, where the annotator must understand the other's distinct state (decoupling) while simultaneously allowing for an emotional connection to it (coupling) (Frith & Frith, 2006; Decety & Jackson, 2004). This dual requirement positions empathy annotation squarely in the middle of our cognitive-affective spectrum .

Appendix N: Robustness to task complexity

To ensure that our main findings were not confounded by variations in complexity between the different annotation tasks (i.e., resilience, empathy, and polarity), we conducted an additional robustness analysis. The objective of this analysis was to isolate the effects of cognitive ability and emotional intelligence on annotation accuracy while explicitly controlling for both task-level and instance-level difficulty.

The analysis involved estimating a pooled regression model with additional controls for complexity. The specific steps were as follows:

1. **Model Specification:** All observations were combined to estimate a model in which annotation accuracy was regressed on participants' cognitive ability and emotional intelligence. Each of these predictors was interacted with a three-level dummy variable representing the task (resilience, empathy, or polarity).
2. **Complexity Controls:** Two layers of controls were incorporated into the model:
    - **Between-Task Fixed Effects:** Task-level intercepts were included to absorb systematic differences in difficulty across the three constructs.
    - **Instance-Level Hardness:** An item-specific hardness index, generated using the Python package PyHard, was included to account for residual difficulty variation within each task.
3. **Technical Details:** Due to model constraints (marginal-effect decomposition is incompatible with both sample fixed effects and cross-task interactions), the individual text sample fixed effects (σ_i) were omitted from this pooled model. Standard errors were clustered at the annotator–task level.
4. **Robustness Checks:** The primary model was checked against three alternative specifications, each maintaining the same covariate structure:
    - A cluster-bootstrap logit model.
    - A random-intercept mixed logit model.
    - A binomial Generalised Estimating Equation (GEE) with an exchangeable working correlation.

The results of this analysis, presented in Table N below, confirm the robustness of our main findings. After controlling for task type and instance-level complexity, the core relationships between cognitive ability, emotional intelligence, and annotation accuracy for each task remain consistent with those reported in Table 6 of the main manuscript.

The only minor deviation observed was a small, statistically significant effect of cognitive ability on polarity identification. This effect is not present in our primary model in the manuscript, which includes text sample fixed effects, suggesting it is accounted for by item-level variance. Overall, this analysis provides strong evidence that our main conclusions are robust to considerations of between-task complexity.

Table N. Results of robustness analysis controlling for task complexity with instance-level hardness and interaction of cognitive ability and emotional intelligence with task type

Task	Δ Pr(correct) per unit ↑ Cognitive ability	Δ Pr(correct) per unit ↑ Emotional intelligence	Take-away
Resilience	**+0.0053*** **	–0.0133 (non-sig)	Only cognitive ability matters; EI has no reliable effect – consistent with our main results.
Empathy	**+0.0088*** **	**–0.0790*** **	Those with higher cognitive ability annotate empathy items more accurately, but higher EI backfires – consistent with our main results.
Polarity	+0.0027** (non-sig in main results after accounting for sample item FEs)	**+0.0437*** **	Accuracy is improved mainly by EI, consistent with our main results. Cognitive ability adds only a small increase, which is eliminated in the model presented in the paper after accounting for fixed effects of the text samples.

***p<0.05, **p<0.01, ***p<0.001; Bolding represents significant results agreeing with those presented in manuscript**

Appendix O: Affective nature of financial news sentiment

Despite the fact that domain-specific information could be utilized to cognitively estimate sentiment in financial news, we argue that evaluating financial sentiment is primarily an affective task.

Dating back to some of the earliest work in this area, a number of studies have directly addressed whether sentiment in financial news has an informational or emotional influence on investor views. Tetlock (2007) directly evaluated these competing theories, noting that information theory predicts that impacts of sentiment

would persist because they represent information about firm fundamentals, but found that instead, shifts in trading in reaction to news sentiment reversed themselves, indicating that this type of news elicits an emotional, not rational reaction. Vargas-Sierra and Orts (2023) demonstrate that even specialized financial newspapers systematically embed emotional lexical cues such as fear, greed, and optimism to guide readers' affective interpretations. It has also been shown that even non-verbal affective cues (which would be more difficult to characterize as activating cognitive pathways) convey significant information in financial contexts (Mayew and Venkatachalam 2012). These are exemplar studies, but broad literature shows affect-based lexicons capture market sentiment, which in turn correlates with stock performance. Collectively, these studies suggest that affective appraisal is integral to how investors interpret financial news.

From an empirical standpoint, our analysis bears out this result. As noted in Table 6 in the manuscript, we find in our pre-registered experiment that emotional intelligence of annotators has a positive impact on their ability to identify sentiment in excerpts from financial news, and that cognitive ability has no such association (despite cognitive ability having a positive impact on the ability to identify constructs of resilience and empathy).

The dataset used for our study is comprised of sentences extracted from financial news in the popular press. As such, the sample items in our corpus largely do not resemble the example provided by the reviewer, and from a face validity standpoint are much more interpretable based solely on affective cues. Some examples:

- Sony Ericsson and Nokia dominated the list of best-selling handsets with five models each
- The company is well positioned in Brazil and Uruguay
- The Dutch broker noted that Nokian Tyres reported a good first quarter in 2006, above or in line with consensus
- Ford is struggling in the face of slowing truck and SUV sales and a surfeit of up-to-date, gotta-have cars
- The airline estimated that the cancellation of its flights due to the closure of European airspace, and the process of recommencing traffic, have caused a the company a loss of EUR20m, including the costs of stranded passengers' accommodation.
- Due to the rapid decrease in net sales, personnel reductions have been carried out on a wider scale than initially expected

In order to further support the claim that sentiment signals in the dataset used for our case study are affective and not reliant on cognitive pathways for interpreting complex financial concepts, we conducted an additional robustness analysis. Specifically, we compared the results obtained when using finance domain-specific versus generic affective lexicons. We first identified all words in the Loughran–McDonald (L&M) lexicon (finance domain-specific) that do not appear in the generic affective NRC Lex (positive/negative) lexicon and then extracted all samples that included at least one of these words. Next, we identified all words in NRC Lex that are not present in L&M and extracted the corresponding samples containing at least one of those words. We reran our main analysis separately for these two subsets.

First, domain specific language was much rarer than general affective language. Of our samples, only 9% contained at least one L&M word but no NRC word, whereas 35% contained least one NRC word but no L&M word (29% contained both). We also re-ran our analysis reported in manuscript Table 6, noting for the subset containing L&M-only words, neither emotional intelligence nor cognitive ability had a statistically significant effect on annotation performance. In contrast, for the subset containing NRC-only words, emotional intelligence remained significant and positive, whereas cognitive ability was not significant. These results held when down-sampling the number of observations for the NRC only dataset to match that of the L&M only dataset. This analysis supports that affective signals are really what is driving annotation of sentiment in our case study, regardless of the specificity of the domain (finance).

Given support from a) prior studies that sentiment in financial news is grounded in affective processes, b) the empirical results in our manuscript that bear this out, c) the prevalence of generic affective content (as opposed to domain-specific jargon) in our corpus, and d) the additional analyses performed to demonstrate that generic affective content is driving annotation performance, we feel confident that our case study is representative of a task aligned with affective abilities.

Table O: Polarity annotation performance breakdown by sample item type			
	L&M no NRC	**NRC no LM**	**NRC no LM (down-sampled)**
Cog_Ability	0.04 (0.03)	0.02 (0.02)	0.05 (0.04)
Emo_Intel	0.47 (0.47)	0.37 * (0.27)	0.68 * (0.33)
Pos_Mood	-0.59 (0.49)	-0.16 (0.14)	-0.33 (0.27)
Neg_Mood	0.33 (0.40)	0.12 (0.14)	0.19 (0.32)
edu_category	-0.16 (0.43)	-0.01 (0.15)	-0.64 (0.37)
age_category	-0.07 (0.27)	0.12 (0.08)	-0.04 (0.17)
Gender & Race Categories	✓	✓	✓
Observations	128	643	120
Sample-fixed Effects	YES	YES	YES

Notes: We estimated conditional (fixed-effects) logistic regressions grouped by unique_id (similar to Table 6 in the manuscript). In conditional logit, groups with no within-group variation in the dependent variable do not contribute to identification and are automatically omitted.
* denotes significance at 0.05.

References:

Abbasi A, Dobolyi D, Lalor JP, Netemeyer RG, Smith K, Yang Y (2021) Constructing a Psychometric Testbed for Fair Natural Language Processing. Moens MF, Huang X, Specia L, Yih SW tau, eds. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. (Association for Computational Linguistics, Online and Punta Cana, Dominican Republic), 3748–3758.

Batson D, Ahmad N, Lishner D, Tsang J, Snyder CR, Lopez SJ (2002) Empathy and altruism. *The Oxford handbook of hypo-egoic phenomena*. 161–174.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, et al. (2020) Language Models are Few-Shot Learners. *arXiv.org*. Retrieved (June 21, 2023), https://arxiv.org/abs/2005.14165v4.

Buechel S, Hahn U (2017) EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. *the Association for Computational Linguistics* 2:578–585.

Buzzanell PM (2010) Resilience: Talking, Resisting, and Imagining New Normalcies Into Being. *Journal of Communication* 60(1):1–14.

Chan D (2010) So Why Ask Me? Are Self-Report Data Really That Bad? *Statistical and Methodological Myths and Urban Legends*. (Routledge).

Chia YK, Chen G, Tuan LA, Poria S, Bing L (2023) Contrastive Chain-of-Thought Prompting. (November 15) http://arxiv.org/abs/2311.09277.

Clark M, Robertson M, Young S (2019) "I feel your pain": A critical review of organizational research on empathy. *Journal of Organizational Behavior* 40(2):166–192.

Clore GL, Huntsinger JR (2007) How emotions inform judgment and regulate thought. *Trends in Cognitive Sciences* 11(9):393–399.
Cuff B, Brown S, Taylor L, Howat D (2016) Empathy: A Review of the Concept. *Emotion Review* 8(2):144–153.
Dajani DR, Uddin LQ (2015) Demystifying cognitive flexibility: Implications for clinical and developmental neuroscience. *Trends Neurosci* 38(9):571–578.
Davidson RJ, Sutton SK (1995) Affective neuroscience: the emergence of a discipline. *Current Opinion in Neurobiology* 5(2):217–224.
Decety J, Jackson PL (2004) The functional architecture of human empathy. *Behav Cogn Neurosci Rev* 3(2):71–100.
De Neys W (2018) *Dual process theory 2.0.* (Routledge/Taylor & Francis Group, New York, NY, US).
Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*4171–4186. (May 24) http://arxiv.org/abs/1810.04805.
Evans JStBT, Stanovich KE (2013) Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspect Psychol Sci* 8(3):223–241.
Evans JSBT (2008) Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol* 59:255–278.
Evans JStBT (2003) In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7(10):454–459.
Finucane ML, Alhakami A, Slovic P, Johnson SM (2000) The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making* 13(1):1–17.
Fiori M (2009) A new look at emotional intelligence: A dual-process framework. *Personality and Social Psychology Review* 13(1):21–44.
Folkman S, Lazarus RS, Dunkel-Schetter C, DeLongis A, Gruen RJ Dynamics of a Stressful Encounter: Cognitive Appraisal, Coping, and Encounter Outcomes.
Frith CD, Frith U (2006) The neural basis of mentalizing. *Neuron* 50(4):531–534.
Hatfield E, Cacioppo JT, Rapson RL (1994) *Emotional Contagion* (Cambridge University Press).
Hickmann ML, Wurzberger F, Hoxhalli M, Lochner A, Töllich J, Scherp A (2022) Analysis of GraphSum's Attention Weights to Improve the Explainability of Multi-Document Summarization. (December 6) http://arxiv.org/abs/2105.11908.
Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Computation* 9(8):1735–1780.
Kahneman D (2011) *Thinking, fast and slow* (Farrar, Straus and Giroux, New York, NY, US).
Kern ML, Park G, Eichstaedt JC, Schwartz HA, Sap M, Smith LK, Ungar LH (2016) Gaining insights from social media language: Methodologies and challenges. *Psychological Methods* 21(4):507–525.
König A, Graf-Vlachy L, Bundy J, Little LM (2020) A Blessing and a Curse: How CEOs' Trait Empathy Affects Their Management of Organizational Crises. *AMR* 45(1):130–153.
Kosinski M, Wang Y, Lakkaraju H, Leskovec J (2016) Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods* 21(4):493–506.
Lazarus RS, Folkman S (1984) *Stress, Appraisal, and Coping* (Springer Publishing Company).
Loughran T, McDonald B (2011) When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance* 66(1):35–65.
Madaan A, Tandon N, Gupta P, Hallinan S, Gao L, Wiegreffe S, Alon U, et al. (2023) Self-Refine: Iterative Refinement with Self-Feedback. (May 25) http://arxiv.org/abs/2303.17651.
Malo P, Sinha A, Takala P, Korhonen P, Wallenius J (2013) Good Debt or Bad Debt: Detecting Semantic Orientations in Economic Texts. (July 23) http://arxiv.org/abs/1307.5336.
Masten AS (2001) Ordinary magic. Resilience processes in development. *Am Psychol* 56(3):227–238.
Mayew, W. J., and Venkatachalam, M. (2012). "The Power of Voice: Managerial Affective States and Future Firm Performance." *Journal of Finance* (67:1), pp. 1–43.
Metcalfe J, Mischel W (1999) A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review* 106(1):3–19.

Mischel W, Ayduk O (2004) Willpower in a cognitive-affective processing system: The dynamics of delay of gratification. *Handbook of self-regulation: Research, theory, and applications* (The Guilford Press, New York, NY, US), 99–129.

Mohebbi H, Zuidema W, Chrupała G, Alishahi A (2023) Quantifying Context Mixing in Transformers. (February 8) http://arxiv.org/abs/2301.12971.

Mousavi R, Gu B (2024) Resilience Messaging: The Effect of Governors' Social Media Communications on Community Compliance During a Public Health Crisis. *Information Systems Research* 35(2):505-527.

Panksepp J (2003) At the interface of the affective, behavioral, and cognitive neurosciences: Decoding the emotional feelings of the brain. *Brain and Cognition* 52(1):4–14.

Pennebaker JW, Francis ME (1996) Cognitive, Emotional, and Language Processes in Disclosure. *Cognition and Emotion* 10(6):601–626.

Pennebaker JW, Boyd RL, Jordan K, Blackburn K (2015) The Development and Psychometric Properties of LIWC2015.

Rheault L (2016) Expressions of Anxiety in Political Texts. Bamman D, Doğruöz AS, Eisenstein J, Hovy D, Jurgens D, O'Connor B, Oh A, Tsur O, Volkova S, eds. *Proceedings of the First Workshop on NLP and Computational Social Science*. (Association for Computational Linguistics, Austin, Texas), 92–101.

Richardson GE (2002) The metatheory of resilience and resiliency. *Journal of Clinical Psychology* 58(3):307–321.

Riepenhausen A, Wackerhagen C, Reppmann ZC, Deter HC, Kalisch R, Veer IM, Walter H (2022) Positive Cognitive Reappraisal in Stress Resilience, Mental Health, and Well-Being: A Comprehensive Systematic Review. *Emotion Review* 14(4):310–331.

Salovey P, Mayer J (1990) Emotional Intelligence Imagination, cognition, and personality. *Imagination, Cognition and Personality* 9(3):1989–90.

Sedoc J, Buechel S, Nachmany Y, Buffone A, Ungar L (2019) Learning Word Ratings for Empathy and Distress from Document-Level User Responses. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*:1664–1673.

Sergent K, Stajkovic AD (2020) Women's leadership is associated with fewer deaths during the COVID-19 crisis: Quantitative and qualitative analyses of United States governors. *Journal of Applied Psychology* 105(8):771–783.

Singer T, Lamm C (2009) The social neuroscience of empathy. *Ann N Y Acad Sci* 1156:81–96.

Shipstead Z, Harrison TL, Engle RW (2016) Working Memory Capacity and Fluid Intelligence: Maintenance and Disengagement. *Perspect Psychol Sci* 11(6):771–799.

Slovic P, Finucane ML, Peters E, MacGregor DG (2007) The affect heuristic. *European Journal of Operational Research* 177(3):1333–1352.

Stanovich KE, West RF (1998) Individual differences in rational thought. *Journal of Experimental Psychology: General* 127(2):161–188.

Suslow T, Hoepfel D, Günther V, Kersting A, Bodenschatz CM (2022) Positive attentional bias mediates the relationship between trait emotional intelligence and trait affect. *Sci Rep* 12(1):20733.

Tausczik YR, Pennebaker JW (2010) The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29(1):24–54.

Tetlock PC (2007) Giving Content to Investor Sentiment: The Role of Media in the Stock Market. *The Journal of Finance* 62(3):1139–1168.

Tetlock PC, Saar-Tsechansky M, Macskassy S (2008) More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance* 63(3):1437–1467.

Vargas-Sierra, C., and Orts, M. Á. (2023). "Sentiment and Emotion in Financial Journalism." *Humanities and Social Sciences Communications* (10:1), Article 69.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is All you Need. *Advances in Neural Information Processing Systems* (Curran Associates, Inc.).

Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2023) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. (January 10) http://arxiv.org/abs/2201.11903.

Wispé L (1986) The distinction between sympathy and empathy: To call forth a concept, a word is needed. *Journal of Personality and Social Psychology* 50(2):314–321.

Wu Z, Nguyen TS, Ong DC (2020) Structured Self-Attention Weights Encode Semantics in Sentiment Analysis. (October 10) http://arxiv.org/abs/2010.04922.

Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of Thoughts: Deliberate Problem Solving with Large Language Models. (December 3) http://arxiv.org/abs/2305.10601.

Yao ZF, Hsieh S (2019) Neurocognitive Mechanism of Human Resilience: A Conceptual Framework and Empirical Review. *Int J Environ Res Public Health* 16(24):5123.

Zaki J, Ochsner KN (2012) The neuroscience of empathy: progress, pitfalls and promise. *Nat Neurosci* 15(5):675–680.

Zhang C, Zou L, Luo D, Tang M, Luo X, Li Z, Li C (2024) Efficient Sparse Attention needs Adaptive Token Release. (arXiv).