In Generative AI We (Dis)Trust? Computational Analysis of Trust and Distrust in Reddit Discussions

Aria Pessianzadeh, Naima Sultana, Hildegarde Van den Bulck, David Gefen, Shahin Jabbari, Rezvaneh Rezapour
Drexel University
{ap3943, shahin, shadi.rezapour}@drexel.edu

Abstract

The rise of generative AI (GenAI) has impacted many aspects of life. As these systems become embedded in everyday practices, understanding public trust in them also becomes essential for responsible adoption and governance. Prior work on trust in AI has largely drawn from psychology and human–computer interaction, but there is a lack of computational, large-scale, and longitudinal approaches to measuring trust and distrust in GenAI and large language models. We present the first computational study of *Trust* and *Distrust* in GenAI, using a multi-year Reddit dataset (2022–2025) spanning 39 subreddits and 230,576 posts. Crowd-sourced annotations of a representative sample were combined with classification models for scale analysis. Our results show that *Trust* and *Distrust* are nearly balanced over time, although *Trust* modestly outweighs *Distrust*. Technical performance and usability dominate as dimensions for both categories, while personal experience is the most frequent reason shaping attitudes. Distinct patterns also emerge across trustor groups: while industry professionals and tech leaders predominantly express *Trust* in GenAI, *Distrust* remains more prevalent among AI ethicists, journalists, and the general public. Our results provide a methodological framework for large-scale *Trust* analysis and insights into evolving public perceptions towards GenAI.

1 Introduction

Generative artificial intelligence (GenAI) and large language models (LLMs) are transforming how people search for information, create content, and interact online (Ferrara, 2024; Liang et al., 2025). Their rapid adoption in classrooms and workplaces (Ghimire and Edwards, 2024; Pettersson et al., 2024), as well as in creative industries (Amato et al., 2019), has triggered global debates about the potential benefits and risks of these technologies (Bender et al., 2021; Crawford, 2021). At the center of these debates lies the question of trust: will the public perceive GenAI as reliable, competent, and aligned with their values, or as opaque, biased, and potentially harmful?

Trust is widely recognized as a prerequisite for adoption, appropriate use, and governance of AI systems (Mayer et al., 1995; Hancock et al., 2020; Lee and See, 2004; Hoff and Bashir, 2015). Public attention to GenAI, and especially to questions of trust and reliability, has surged in parallel with the release of advanced LLMs. The launch of ChatGPT in late 2022, followed by successive breakthroughs such as GPT-4 and LLaMA2, amplified both enthusiasm and concern, making trust a focal point of GenAI discourse (Shen et al., 2023; Weidinger et al., 2022). Understanding how trust and distrust are expressed in response to such advancements is crucial for capturing the dynamics of public attitudes and informing the responsible design and governance of GenAI.

Prior research on trust in GenAI has primarily focused on theoretical frameworks or small-scale qualitative studies (Schuetz et al., 2025; Balayn et al., 2024). Recent computational work has analyzed public sentiment toward LLMs using approaches such as Critical Discourse Analysis (Heaton et al., 2024; Emdad et al., 2023), but focused on domain-specific contexts like education or health care rather than systematically mapping diverse dimensions of trust and distrust. Attitudes have also been examined through validated scales such as the Attitudes Toward General LLMs (ATGLLM) and Attitudes Toward Primary LLMs (ATPLLM) (Liebherr et al., 2025), while TAM (Technology Acceptance Model)-based studies showed that trust, privacy, and misuse concerns strongly shape AI adoption (Abdaljaleel et al., 2023; Kelly et al., 2025). Although prior work has developed valuable conceptual frameworks for trust in

Transactions of the Association for Computational Linguistics, vol. 14, pp. 1633–1659, 2026. https://doi.org/10.1162/TACL.a.744
Action Editor: Yulan He. Submission batch: 10/2025; Revision batch: 2/2026; Published 7/2026.

GenAI and LLMs, there is limited large-scale, longitudinal, computational analysis of how public attitudes toward GenAI evolve.

We address this gap with a multi-year analysis of Reddit discussions (Nov 2022–Jun 2025). Reddit offers a suitable corpus given its scale, topical diversity, and sustained discourse around model releases. We collected 230,576 GenAI-related posts from 39 subreddits and crowd-sourced annotations on a representative sample to capture *Trust*, *Distrust*, their dimensions, reasons, and trustors, then scaled these analyses using transformer-based and state-of-the-art LLMs. Our findings show that *Trust* moderately outweighs *Distrust* across the timeline, with shifts around major releases. At the dimension level, *Competence*, *Reliability*, and *Familiarity* dominate, underscoring functionality and personal use as key drivers. Personal experience emerges as the most salient reason for *Distrust* and *Trust*, consistent with recent industry reports (OpenAI, 2025; Anthropic, 2025). Finally, different trustors (e.g., experts, ethicists, business leaders) exhibit distinct patterns, highlighting the sociotechnical complexity of GenAI discourse. For example, industry professionals and tech leaders show greater trust in GenAI, whereas distrust is more prevalent among AI ethicists, journalists, and the general public.

This work contributes the first large-scale dataset of Reddit posts on GenAI labeled for *Trust*, *Distrust*, dimensions, reasons, and trustors; a methodological framework for classifying *Trust* and *Distrust* in social media discourse; and offers new empirical insights into how *Trust* and *Distrust* toward GenAI evolve, what dimensions dominate, and how different actors express them. By combining computational analysis with sociotechnical framing, our study advances the understanding of trust in GenAI and provides tools for monitoring and contextualizing public perceptions. Our data (Pessianzadeh et al., 2026) is available at `https://zenodo.org/records/20331918` and our code can be found at `https://github.com/social-nlp-lab/trust-GenAI`.

2 Related Work

Trust in Technology & AI Systems. Foundational research on trust in technology provides key perspectives on how trust is formed. Trust in the social sciences is defined as the willingness to be vulnerable by relying on another, setting aside concerns about potential negative behavior (Mayer et al., 1995; Luhmann, 2018). Schuetz et al. (2025) found perceptions and experiences with technology, trust transfer, and calculative trust as three categories of influential factors that shape trust. Similarly, Lee and See (2004) highlighted how context, system characteristics, and cognitive processes influence trust relationships between users and automated systems. Glikson and Woolley (2020) showed that cognitive trust stems from transparency and reliability, while emotional trust arises through anthropomorphism. Jacovi et al. (2021) formalized AI–human trust as a cognitive model grounded in user vulnerability and anticipation of AI decisions, where trust depends on expectations that agreements will hold. Balayn et al. (2024) emphasized the complexity of LLM supply chains, where trust dynamics extend from foundational model developers to downstream companies and end users, influenced by both personal propensity and context of use. In parallel, Kim et al. (2023) discussed the interplay of human factors (e.g., domain knowledge), AI factors (e.g., ability, integrity, benevolence), and contextual conditions (e.g., task difficulty, perceived risks) in determining levels of trust in AI systems. These mechanisms also apply in domains such as electronic marketplaces (Pavlou and Gefen, 2004). Importantly, distrust is not defined as the absence of trust but distinct expectations of negative behavior (Lewicki et al., 1998), supported by evidence that trust and distrust rely on different neurocognitive processes (Dimoka, 2010). Trust generally increases behavioral intentions, whereas distrust reduces them (Gefen et al., 2025).

Computational Analysis of Trust. Recent work has developed computational and survey-based measures of trust. The Attitudes Toward General LLMs and Attitudes Toward Primary LLMs scales were developed and validated for assessing acceptance and fear in LLM contexts, distinguishing between societal perceptions and individual user experiences (Liebherr et al., 2025). Computational analyses of public discourse also has revealed mixed sentiment: topic and sentiment modeling of Reddit discussions about ChatGPT and DeepSeek found recurring concerns around bias, ethical risks, and transparency (Katta, 2025). Building on the TAM, Abdaljaleel et al. (2023) introduced the TAME-ChatGPT instrument to study

healthcare students' attitudes, while Kelly et al. (2025) found that trust and privacy concerns differentially influence ChatGPT acceptance in mental versus physical healthcare.

LLMs and Trust. Research specific to LLMs highlights the multi-layered nature of trust and distrust. Paraschou et al. (2025) proposed the 'bowtie model' to conceptualize trustor–trustee relationships, showing how perceptions of competence, knowledge, and transparency shape user behavior. Davoodi and Mezei (2024) applied computational methods to customer feedback and identified delivery processes as central to trust formation. Broader reviews highlight conceptual fragmentation: Mehrotra et al. (2024) stressed definitional inconsistencies, while Sousa et al. (2023) argued that trust is often oversimplified. Gerlich (2024) found that many individuals trust AI more than humans, attributing this to their perceived impartiality, but other studies found otherwise (Gefen et al., 2025), suggesting that trust might be context-dependent. Recent mapping studies confirm rising attention to LLM trustworthiness, with transparency, explainability, and reliability as recurring themes (de Cerqueira et al., 2025). Computational studies further analyze public opinion toward LLMs across social platforms. Heaton et al. (2024) showed that Twitter users often frame ChatGPT as a social actor, while Emdad et al. (2023) and Takagi (2023) examined Reddit discussions on ChatGPT in education, finding mostly neutral or contested views. Skjuve et al. (2023) emphasized pragmatic and hedonic features in shaping trust. Other work used machine learning for trust modeling (Zapata et al., 2023; Yahyaoui and Al-Mutairi, 2016; Wang et al., 2020; López and Maag, 2015; Hauke et al., 2013), while Gefen et al. (2024) demonstrated that perceptions of trust in GenAI evolve from technical concerns toward broader societal issues. Prior work provides valuable conceptual and computational insights but remains fragmented, domain-specific, and small in scale. Few studies track how trust and distrust in GenAI evolve in response to major model releases, a gap this study addresses.

3 Method

3.1 Data Collection

To examine narratives surrounding GenAI on Reddit following the release of ChatGPT (November 30, 2022), we implemented a two-step data collection strategy. We first identified relevant subreddits and then applied content-based filtering to isolate GenAI-related posts.

Step 1: Identification of Subreddits. We systematically identified communities focused on LLMs, their developers, AI more broadly, and related technologies. We reviewed approximately 90 candidate subreddits, examining their content, "About" pages, topical focus, and creation dates. Based on relevance and sustained activity, we selected 39 subreddits that were active between November 30, 2022, and June 30, 2025. Using Pushshift (Baumgartner et al., 2020) and Arctic-shift (Heitmann, 2025), we retrieved 1,944,899 posts from these subreddits.

Step 2: Identification of GenAI-Related Posts. To identify posts specifically related to GenAI within this corpus, we employed a mixed-method approach combining keyword-based filtering and topic modeling. We first compiled a list of seed terms related to GenAI and expanded it to include the names of major models (e.g., prominent LLMs), as well as key concepts and topics associated with generative AI. To further refine and expand this list, we applied Latent Dirichlet Allocation (LDA) (Blei and Lafferty, 2009) to posts from the two most active subreddits (*r/ChatGPT* and *r/OpenAI*) and two broader AI-focused subreddits (*r/ArtificialIntelligence* and *r/MachineLearning*). We examined the resulting topics to identify additional relevant terms and incorporated them into our keyword list. Using this comprehensive set of GenAI-related keywords, we retained posts containing at least one matching term (see Appendix A for the list of subreddits and keywords).

We excluded posts in which the author or content was marked as "removed" or "deleted" (to protect user privacy), posts with empty bodies, and posts containing fewer than five words. After applying these filtering criteria, the final dataset comprised 230,575 posts.

3.2 Conceptual Frameworks

3.2.1 Dimensions of Trust and Distrust

Trust and distrust are central but distinct constructs in human–AI interaction. Drawing from psychology, sociology, and HCI research, we define *Trust* as the willingness to rely on a system with positive expectations of its reliability, competence, and intentions, even under conditions of uncertainty or vulnerability (Mayer et al., 1995; Rousseau et al.,

	Dimension	Definition
Trust	Reliability	Information accuracy and consistency
	Transparency	Openness about GenAI processes and data use
	Familiarity	Previous exposure to GenAI or similar systems
	Integrity	Adherence to ethical principles acceptable to users
	Competence	Effective functionality and features
	Benevolence	Belief that GenAI acts with good intentions
Distrust	Unreliability	Lack of accuracy and consistency
	Opaqueness	Lack of transparency in data use
	Unfamiliarity	New or unfamiliar technology causing doubt
	Dishonesty	Failure to follow user-accepted ethical principles
	Incompetence	Inability to perform tasks effectively
	Deception	Perception of dishonesty or hallucination in outputs
	Malevolence	Belief that GenAI has harmful or malicious intent

Table 1: Dimensions of *Trust* and *Distrust*

1998). In contrast, *Distrust* refers to active suspicion and negative expectations, grounded in concerns about unreliability, deception, or potential harm (Lewicki et al., 1998; Harrison McKnight and Chervany, 2001). Importantly, *Distrust* is not merely the absence of *Trust*; rather, it is a qualitatively distinct phenomenon with its own logic and consequences (Lewicki et al., 1998).

Grounded in extensive literature review of trust in technology and AI (Zhang et al., 2023; McKnight et al., 2009; Bach et al., 2024; Braga et al., 2018; Xu et al., 2014; Lankton et al., 2015; Zhang et al., 2024; Kim et al., 2024; Klingbeil et al., 2024; Salimzadeh et al., 2024; Shin, 2021; Blöbaum et al., 2016; McKnight et al., 2017; Horowitz et al., 2024; Schuetz et al., 2025), we first compiled 37 candidate dimensions of *Trust* and *Distrust* (see Appendix B). A team of four annotators manually coded a sample of 100 posts to evaluate the relevance and applicability of these dimensions in the context of GenAI. Through iterative rounds of coding and discussion, we removed overlapping concepts and distilled the list into 6 key dimensions of *Trust* (*Reliability, Competence, Familiarity, Integrity, Transparency, Benevolence*) and 7 key dimensions of *Distrust* (*Unreliability, Incompetence, Unfamiliarity, Deception, Dishonesty, Opaqueness, Malevolence*), adapted to LLMs and GenAI (see Table 1).

3.2.2 Trustor and Reason of Trust

We further identified two additional concepts: (1) the *trustor*, i.e., who expresses trust or distrust, and (2) the underlying *reason* for that judgment.

Trustor. In trust relationships, the *trustor* is the entity placing confidence or suspicion in a system, while the *trustee* is the system or actor being judged (Mayer et al., 1995; Paraschou et al., 2025). In GenAI, the trustor may range from end-users to institutions shaping AI narratives. Their identity influences both expectations and evaluation criteria (Balayn et al., 2024; Kim et al., 2023); for example, developers may emphasize transparency and reliability, while ethicists may focus on fairness and accountability. To capture this diversity, we defined 10 categories through inductive coding: GenAI users, software developers, researchers or academics, tech industry professionals, general public, media and journalists, business leaders or executives, AI ethicists or advocacy groups, artists or creatives, and educators or knowledge workers.

Reason. The *reason* for trust or distrust reflects the evidence upon which the trustor bases their judgment. Social psychology and HCI research show that reasons stem from prior experience, social influence, or perceived organizational responsibility (Lee and See, 2004; Glikson and Woolley, 2020). In online discourse, these may be direct (e.g., personal use) or indirect (e.g., media or peer reports). We defined 6 categories: personal experience using AI models, reports from media or experts, discussions on social media or forums, general perceptions of AI technology, general perceptions of responsible companies, and experiences shared by friends, family, or colleagues.

3.3 Crowdsourcing and Data Annotation

We developed a codebook with definitions and examples of *Trust*, *Distrust*, and their dimensions to guide the annotation process. We opted for crowdsourcing rather than expert-only annotation to leverage the broader demographic diversity and scalability it provides (Sap et al., 2021; Wan et al., 2023). The annotation task was hosted on Potato (Pei et al., 2022) and deployed through Prolific. After multiple rounds of testing, annotators received the detailed guidelines and were tasked to select: (1) if a post expressed *Trust*, *Distrust*, *Both*, or *Neither* toward GenAI; (2) relevant dimensions of *Trust* or *Distrust*; (3) the *trustor* (the entity expressing *Trust* or *Distrust*); and (4) underlying *reason* from our predefined list.

Each post was independently annotated by five annotators to ensure labeling reliability. Across three annotation rounds, annotators labeled $2,690$ posts. For the first task, we only retained posts for which at least three out of five annotators agreed on the label (choosing the majority vote); posts

Trust	*I love having ChatGPT explain technical aspects to me as if I were a medieval noble receiving counsel from a trusted advisor. My life becomes so much less mundane.*
	Currently using ChatGPT to create a cheat sheet for units in my physics. This is 100 times easier to learn as well. Did I just make my teacher's job obsolete?
Distrust	*GPT4 is working really badly all of a sudden. Anyone else noticing this? It takes forever, and fails to respond with an error.*
	Why GPT4 just no longer follow instructions? Is it just for me, or is GPT4 no longer following instructions? Simple prompts that have always worked are now suddenly misinterpreted by ChatGPT.
Both	*The accuracy of the responses generated by ChatGPT depends on many factors, e.g., quality of the input and complexity of the task. Generally, ChatGPT has shown to produce highly coherent and human-like responses, especially for tasks that involve NLP, language translation, and generating textual content. However, like any AI model, ChatGPT can sometimes make errors or produce inaccurate responses, particularly where it lacks context or faces new or unexpected situations.*
	What if most answers from ChatGPT are wrong? But we just oversaw, as we are still amazed by the awesomeness of quickly responding to complicated inquiries.

Table 2: Examples of posts labeled as *Trust*, *Distrust*, and *Both*

without majority agreement were discarded. After applying quality control, $2,227$ posts were retained. Inter-annotator reliability was assessed using Fleiss' κ, resulting in $\kappa = 0.42$, with an average agreement rate of 77%. Among the retained posts, 877 (39%) had agreement from three out of five annotators, 800 (36%) had agreement from four out of five annotators, and 550 (25%) achieved unanimous agreement. More details about the annotation procedure, the codebook, and guidelines are in Appendix C. Table 2 presents two example posts for each class.

For dimensions of trust and distrust, we only focused on posts kept from the previous task. For each post, we retained the dimension annotations provided by the annotators whose labels formed the majority in the previous task. To construct ground-truth labels for the dimensions, we applied two aggregation schemes: *Majority*: dimensions selected by more than half of the annotators, and *Any*: dimensions selected by at least one of the annotators. To compute inter-annotator agreement for dimensions, we evaluated trust dimensions only for posts labeled as *Trust* and distrust dimensions only for posts labeled as *Distrust*. The average agreement across the 13 dimensions was 0.82. Similarly, for classifying *Trustor* and *Reason*, the final labels were obtained based on majority vote. Average agreement rate between annotators was 0.72 for *Trustor* and 0.81 for *Reason*.

3.4 Classification

3.4.1 Trust and Distrust

We employed a multi-label classification framework with four labels in task 1: *Trust*, *Distrust*, *Both*, and *Neither*. Our annotated dataset was split into training, validation, and test sets using a 0.7:0.1:0.2 ratio, stratified by label to preserve class distribution. The splits contained $1,558$, 223, and 446 posts for train, validation, and test, respectively.

Transformer-based Models. We fine-tuned five pretrained transformer models, including BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), DistilBERT (Sanh et al., 2019), ModernBERT (Warner et al., 2024), and SocBERT (Guo and Sarker, 2023), on the training set. Fine-tuning was conducted with the `AdamW` optimizer, `batch size` 8, `learning rate` $2e-5$, and up to 3 `epochs` with early stopping based on validation loss. Hyperparameters were selected through validation performance.

LLMs. We evaluated closed-weight models from OpenAI (GPT-5.2-2025-12-11, GPT-5.1-2025-11-13, GPT-5-2025-08-07, GPT-5-mini-2025-08-07, GPT-5-nano-2025-08-07, GPT-4o-2024-11-20, GPT-4o-mini-2024-07-18, GPT-4.1-2025-04-14, GPT-4.1-mini-2025-04-14, GPT-4.1-nano-2025-04-14) (OpenAI, 2024), Claude (claude-sonnet-4-6, claude-haiku-4-5) (Anthropic, 2024), as well as open-weight models such as GPT-oss, Llama3-8B (AI@Meta, 2024), and Mixtral-8x7B (Jiang et al., 2024). For these models, we designed both zero-shot and few-shot prompts (see Appendix D). Few-shot prompts included 3 randomly selected examples per label from the same validation set. We set `temperature` = 0.0.

Evaluation. We evaluated all models on the held-out test set. Weighted F1, per-class precision, and recall were used as metrics.

3.4.2 Dimensions of Trust and Distrust

Building on the previous task, we extended the classification framework to predict the 13 fine-grained dimensions of *Trust* and *Distrust* defined in Table 1. Due to limited annotated data and the

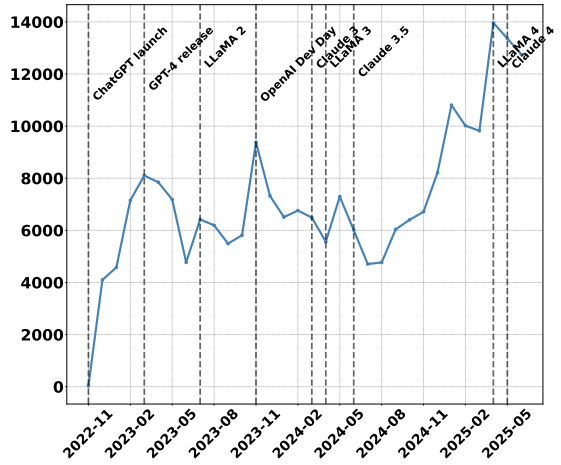


Figure 1: Temporal changes in the number of posts around GenAI after the release of ChatGPT

sparsity of several dimensions, we relied exclusively on LLMs for this task, as the dataset was insufficient to train or fine-tune transformer-based classifiers. For each post, the model was prompted to identify one or more applicable dimensions. To avoid excessively long prompts and to ensure category relevance, the applicable prompt was conditioned on the reported label: if the post was labeled as *Trust* in Task 1, the model was prompted only with trust dimensions; if *Distrust*, only distrust dimensions; and if *Both*, with both sets of dimensions.

Prompting Setup. We designed both zero-shot and few-shot prompts, including 3 illustrative cases per dimension in the few-shot setting. All prompts included the definitions of dimensions from our codebook (see Appendix D). We set `Temperature` = 0.0 and `top-p` = 1.0. For each post, the model could return multiple dimensions.

Evaluation. We evaluate model performance under two ground truth label aggregation schemes defined in § 3.3: *Majority* and *Any*. Treating each dimension as a binary label, we computed precision, recall, and F1 score per dimension, along with macro-averaged scores across all dimensions, following standard multi-label evaluation practices (Madjarov et al., 2012; Tsoumakas and Katakis, 2008).

3.4.3 Trustor and Reason of Trust

LLMs were prompted to assign one trustor and one reason per post, separately, with results evaluated against our test set using precision, recall, and F1 (see Appendix D for prompts).

4 Results

Volume of Discourse on Reddit. Figure 1 shows how the volume of Reddit discourse about GenAI evolved between November 30, 2022, and June 30, 2025. The x-axis represents time (monthly bins), and the y-axis represents the number of posts containing GenAI keywords (after preprocessing). Overall, we observe a gradual yet sustained increase in activity, punctuated by sharp peaks coinciding with major model announcements. The initial surge in late 2022 coincides with the public release of ChatGPT, which brought LLMs into mainstream public attention. Subsequent spikes correspond to milestone releases such as GPT-4 (March 2023), LLaMA2 (July 2023), and OpenAI's Dev Day (November 2023). While discussion levels fluctuated throughout 2023 and early 2024 despite new releases such as Claude3 (March 2024) and LLaMA3 (April 2024), the sharp and sustained growth at the end of 2024 into 2025 culminated in the highest levels of discourse during the releases of models such as LLaMA4 and Claude4 (April-May 2025). Although volume fluctuations may reflect multiple factors, the alignment of peaks with well-publicized model releases may suggest that breakthrough announcements play a key role in driving community engagement.

4.1 Trust and Distrust

Classification Results. Table 3 reports performance across transformers, zero-shot, and few-shot LLMs for classifying *Trust*, *Distrust*, *Both*, and *Neither*. Among transformers, RoBERTa performed the best (F1 = 0.77), with ModernBERT and DistillBERT close behind (F1 = 0.75), reflecting the advantages of more recent pretrained models. Zero-shot LLMs generally outperformed transformers: GPT-5.1, GPT-5-mini, and GPT-5.2 reached F1 = 0.8, 0.79, and 0.79, respectively. Open-source LLMs such as Llama-4-Maverick (F1 = 0.75) lagged. Few-shot LLMs showed a consistent performance, with GPT-5-mini (F1= 0.8), GPT-4.1-mini (F1= 0.79), and GPT-4o (F1= 0.79) leading. The best overall performance was achieved by GPT-5.1 in the zero-shot setting, and GPT-5-mini in the few-shot setting (F1 = 0.8). Our results suggest that adding shots does not necessarily improve performance in this classification task (See Appendix G.3).

Table 4 breaks down GPT-5.1's zero-shot per-

Type	Model	Precision	Recall	F1
Transformers	BERT	0.77	0.75	0.68
	RoBERTa	0.76	0.79	0.77
	DistillBERT	0.74	0.78	0.75
	ModernBERT	0.74	0.76	0.75
	SocBERT	0.73	0.74	0.69
Zero-Shot	GPT-5.2	0.80	0.78	0.79
	GPT-5.1	0.81	0.80	**0.80**
	GPT-5	0.80	0.74	0.76
	GPT-5-mini	0.80	0.79	0.79
	GPT-5-nano	0.79	0.77	0.78
	GPT-4.1	0.79	0.74	0.74
	GPT-4.1-mini	0.79	0.77	0.77
	GPT-4.1-nano	0.77	0.66	0.66
	GPT-4o	0.78	0.74	0.75
	GPT-4o-mini	0.80	0.71	0.73
	GPT-oss-20b	0.79	0.75	0.76
	Claude-sonnet-4-6	0.80	0.72	0.74
	Claude-haiku-4-5	0.79	0.70	0.72
	Mixtral-8x7B	0.80	0.56	0.62
	Llama-4-Maverick	0.76	0.76	0.75
Few-Shot	GPT-5.2	0.81	0.76	0.78
	GPT-5.1	0.79	0.78	0.79
	GPT-5	0.81	0.75	0.77
	GPT-5-mini	0.81	0.79	**0.80**
	GPT-5-nano	0.79	0.77	0.78
	GPT-4.1	0.81	0.75	0.77
	GPT-4.1-mini	0.80	0.78	0.79
	GPT-4.1-nano	0.76	0.61	0.62
	GPT-4o	0.80	0.79	0.79
	GPT-4o-mini	0.81	0.75	0.77
	GPT-oss-20b	0.81	0.76	0.78
	Claude-sonnet-4-6	0.82	0.77	0.78
	Claude-haiku-4-5	0.77	0.72	0.73
	Mixtral-8x7B	0.79	0.37	0.47
	Llama-4-Maverick	0.79	0.74	0.75

Table 3: Model performance for *Trust* vs. *Distrust*

Label	Precision	Recall	F1
Distrust	0.86	0.91	0.89
Trust	0.88	0.80	0.84
Neither	0.55	0.65	0.60
Both	0.50	0.15	0.24

Table 4: Performance of the best model (GPT-5.1 *zero-shot*) in *Trust* vs. *Distrust* classification.

Label	Annotated	All Data	% (All Data)
Neither	369	95,513	41
Trust	812	71,431	31
Distrust	979	60,668	26
Both	67	2962	1

Table 5: Distribution of labeled posts in our data.

formance for each category. Performance was strongest on *Distrust* (F1 = 0.89) and *Trust* (F1 = 0.84), indicating that the model reliably captured clear positive or negative context. Performance dropped for *Neither* (F1 = 0.60), showing difficulty in distinguishing neutral commentary from implicit evaluations. The lowest scores occurred for *Both* (F1 = 0.24), highlighting the challenge in detecting ambivalent or ironic posts (see Appendix G for error analysis). Based on these results and the overall performance of GPT-5.1-2025-11-13 (zero-shot), we used this model to generate labels for the full dataset with OpenAI's batch API.

Label Distribution. Table 5 presents the distribution of labels in the annotated and the full datasets. The largest category is *Neither* (41%), indicating that many GenAI-related posts are descriptive or informational (e.g., sharing news or announcements). *Trust* (31%) and *Distrust* (26%) are nearly balanced, reflecting the polarized nature of public opinion. By contrast, *Both* (1%) is marginal, suggesting that ambivalent expressions are comparatively rare and harder to detect.

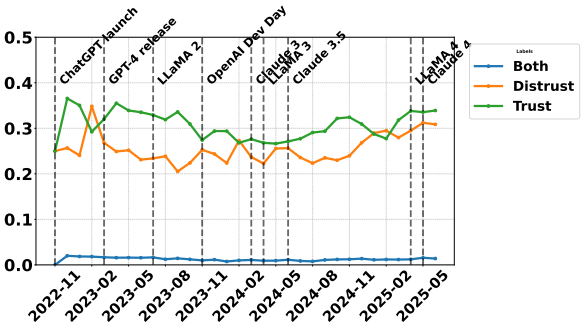


Figure 2: Comparison between *Trust* vs. *Distrust*.

Temporal Changes in Trust & Distrust. Figure 2 shows the monthly changes in the number of posts labeled as *Trust*, *Distrust*, and *Both*, normalized by the total number of posts in that month. The result suggests that *Trust* consistently, but modestly, outweighs *Distrust* across the timeline of our data. While *Trust* accounts for roughly 30–35% of posts across most months, *Distrust* appears in roughly 20-30 % of posts. *Both*, however, accounts for a very insignificant share of posts (We did not add *Neither* to the figure, but it presents the largest number of posts in our data). Despite some fluctuations, neither category extremely dominates the discourse, underscoring a persistent tension in public attitudes toward GenAI (Marantz, 2024; University of Edinburgh Impact, 2023; Hill, 2023). The flat trend for *Both* indicates that ambivalence is rare compared

	2023-02		2024-02		2025-02	
Label	Term	Z-Score	Term	Z-Score	Term	Z-Score
Trust	help	9.55	model	9.28	love	13.68
	use	8.71	AI	8.95	best	13.54
	create	8.03	feedback	8.92	share	10.71
	ChatGPT	7.79	building	8.83	better	10.48
	text	7.23	project	8.65	AI	10.10
Distrust	DAN	-6.13	says	-5.91	weird	-6.00
	wrong	-7.42	refuses	-5.95	deleted	-6.21
	ERP	-7.73	issue	-6.59	worse	-6.50
	Luka	-11.35	asked	-7.10	issue	-8.45
	Replika	-13.21	wrong	-7.23	wrong	-8.65

Table 6: Z-scores of terms associated with *Trust* and *Distrust* during periods when the two concepts converged, based on Figure 2.

to strong positive or negative stances.

Based on the results shown in Figure 2, we observe that *Distrust* slightly exceeded *Trust* during two periods (2023-02 and 2025-02), while the two concepts converged in 2024-02. To further examine the discourse during these three months, we conducted separate Z-score analyses (Monroe et al., 2008) for each period, identifying the most distinctive aspects associated with *Trust* and *Distrust*. Table 6 shows that, in 2023-02, the most distinctive aspects of conversations related to *Distrust* include terms such as *Replika*, *Luka* (its parent company), and *DAN* (Do Anything Now), reflecting ethical concerns surrounding companionship models and unrestrained AI behavior. In contrast, the *Trust* discourse during the same period centered on practical utility through terms such as *help*, *use*, and *create*. In 2024-02 and 2025-02, the distinctive features associated with *Distrust* shifted toward technical frustrations, with terms such as *issue*, *refuses*, and *deleted*. At the same time, *Trust* became increasingly associated with collaborative and affective language, including terms such as *feedback*, *project*, *love*, and *best*. The recurrence of *wrong* in all three periods could indicate that perceived inaccuracy remained a stable source of *Distrust* despite changing broader concerns.

4.2 Dimensions of Trust and Distrust

Classification Results. Table 7 shows the selected results of the dimension classification (full result in Table F1 in Appendix). Under the *majority* condition, using zero-shot promoting, models achieved their highest scores on functionality-related and explicit dimensions, particularly *Competence*, *Incompetence*, and *Unreliability*, where GPT-5 variants and Llama-4 obtained the highest scores overall. In contrast, interpersonal and normative dimensions such as *Benevolence*, *Integrity*, and especially *Transparency* remained comparatively weak, with low peak scores across all models. In the few-shot setting, improvements were most evident for relational dimensions including *Reliability*, *Dishonesty*, and *Familiarity*, where GPT-4.1 and GPT-5 showed higher and more stable performance. However, even with few-shot prompting, dimensions capturing ethical aspects, e.g., notably *Integrity* and *Transparency*, continued to exhibit limited classification performance relative to other categories. Under the *any* criterion (full result in Table F2 in Appendix), overall scores increased in both prompting settings. In the zero-shot, models again performed best on Competence and *Incompetence*, with several models reaching almost perfect performance, while *Unreliability* also achieved consistently high scores across models. *Reliability* improved substantially under few-shot prompting, where GPT-4.1–series models achieved the highest scores. In contrast, *Deception* and *Dishonesty* were slightly better classified with zero-shot, suggesting that these dimensions are less dependent on additional contextual examples. Dimensions such as *Integrity* and *Transparency* remained comparatively difficult to classify across both techniques, indicating persistent difficulty in detecting more abstract or implicitly expressed dimensions.

Based on the results and comparing recall, precision, and F1 score of the models, we observe that the *majority* approach penalizes over-prediction of dimensions, rewarding consensus. On the other hand, the *any* approach rewards identifying dimensions that are more challenging to detect, encouraging the diversity in how dimensions of trust or trust can be implied. Given the variability in performance across dimensions and the absence of a single model that consistently outperformed the others, we selected the best-performing model for each dimension based on the *any* evaluation results. We then applied this dimension-specific model selection strategy to annotate the full dataset.

Label Distribution. Table 8 reports the frequency of dimensions in the annotated set (for two ground-truth methods) and the entire dataset, predicted by LLMs. Within *Trust*-labeled posts,

Strategy	Type	Model	Dimensions													Average
			Ben	Com	Dec	Dis	Fam	Inc	Int	Mal	Opa	Rel	Tra	Unf	Unr	F1
Majority	Zero-Shot	Llama-4	0.36	**0.89**	0.45	0.41	0.25	0.75	0.22	0.51	0.27	0.43	0.19	0.14	0.74	0.43
		GPT-5-2025-08-07	0.32	0.88	0.43	0.42	0.32	0.74	**0.29**	0.52	0.28	0.42	0.20	0.11	0.67	0.43
		GPT-5-mini-2025-08-07	0.36	**0.89**	0.42	0.45	0.31	**0.76**	0.25	0.55	0.34	0.68	0.20	0.11	0.79	0.47
		GPT-4o-2024-11-20	0.35	**0.89**	0.45	0.39	0.30	0.73	0.20	0.43	**0.39**	0.46	0.14	0.15	0.77	0.43
	Few-Shot	GPT-5.1-2025-11-13	0.38	0.88	0.47	0.36	0.32	0.71	0.19	0.41	0.32	**0.69**	**0.24**	0.17	0.77	0.45
		GPT-5-mini-2025-08-07	0.36	**0.89**	0.45	0.42	**0.34**	0.73	0.25	0.51	0.34	0.57	0.19	0.15	**0.80**	0.46
		GPT-4o-mini-2024-07-18	**0.45**	0.88	0.44	**0.51**	0.30	0.73	0.20	0.44	0.35	0.57	0.15	0.13	0.78	0.46
		GPT-4.1-2025-04-14	0.43	0.88	**0.49**	0.37	**0.34**	0.75	0.23	0.44	0.38	0.67	0.20	0.12	**0.80**	0.47
Any	Zero-Shot	Mixtral-8x7B	0.47	0.95	0.35	0.45	0.70	0.48	0.12	0.33	0.53	0.56	0.25	0.15	**0.93**	0.48
		Llama-4	0.32	**0.97**	0.56	0.42	0.30	0.86	0.23	0.42	0.20	0.41	0.11	0.09	0.75	0.43
		GPT-5-mini-2025-08-07	0.41	**0.97**	0.53	**0.52**	0.73	**0.87**	0.20	0.36	0.52	0.68	0.23	0.23	0.84	0.55
		GPT-4o-2024-11-20	0.43	**0.97**	**0.59**	0.45	0.75	0.77	0.10	0.53	0.44	0.44	0.13	0.20	0.82	0.51
		GPT-4.1-2025-04-14	0.29	0.96	0.52	0.36	**0.77**	0.81	0.10	0.49	0.48	0.58	0.09	0.30	0.85	0.51
	Few-Shot	Mixtral-8x7B	**0.57**	0.95	0.22	0.51	0.68	0.71	**0.26**	0.46	0.55	0.66	0.35	0.23	0.90	0.54
		GPT-5-2025-08-07	0.26	**0.97**	0.40	0.41	0.58	0.78	0.16	0.42	0.45	0.47	0.21	0.18	0.74	0.46
		GPT-5-mini-2025-08-07	0.40	**0.97**	0.49	0.44	0.72	0.72	0.17	0.44	**0.57**	0.56	0.26	0.18	0.83	0.52
		GPT-4.1-2025-04-14	0.39	0.96	0.52	0.36	0.61	0.74	0.14	0.53	0.50	**0.74**	0.12	0.18	0.83	0.51
		GPT-4.1-mini-2025-04-14	0.54	0.92	0.47	0.24	0.62	0.61	0.10	**0.60**	0.38	0.63	0.20	0.16	0.76	0.48

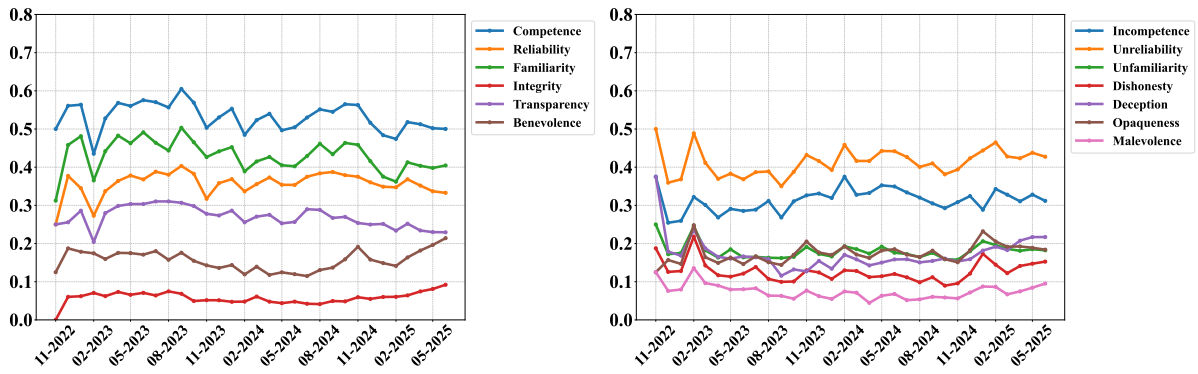
Table 7: Selected classification result of dimensions using zero-shot and few-shot prompting, and any and majority strategies. Dimensions abbreviated as *Benevolence* (Ben), *Competence* (Com), *Deception* (Dec), *Dishonesty* (Dis), *Familiarity* (Fam), *Incompetence* (Inc), *Integrity* (Int), *Malevolence* (Mal), *Opaqueness* (Opa), *Reliability* (Rel), *Transparency* (Tra), *Unfamiliarity* (Unf), *Unreliability* (Unr). Full result in Tables F1 and F2 in Appendix.

	Dimension	Annotated (Majority/Any)	All Data
Trust	Competence	747/ 1095	70,851
	Reliability	611/ 1068	48,011
	Familiarity	176/ 793	57,496
	Transparency	82/ 649	35,514
	Benevolence	61/ 416	21,887
	Integrity	58/ 580	8,480
Distrust	Unreliability	689/ 1179	56,257
	Incompetence	563/ 1081	42,312
	Opaqueness	114/ 700	24,407
	Deception	159/ 676	23,478
	Dishonesty	156/ 709	17,631
	Unfamiliarity	23/ 454	24,469
	Malevolence	60/ 332	10,303

Table 8: Frequency of dimensions in the annotated and full datasets. Each post can have more than one dimension. For the full dataset, we only used *Any* as our labeling strategy.

Competence, *Reliability*, and *Familiarity* were the most common, reflecting how users primarily evaluate GenAI in terms of technical performance and personal experience. By contrast, ethical or normative dimensions such as *Integrity*, *Transparency*, and *Benevolence* appeared less frequently. Within *Distrust*-labeled posts, *Unreliability*, and *Incompetence* dominated, echoing concerns about functionality and accuracy. Other dimensions, such as *Deception*, *Dishonesty*, and *Opaqueness*, occurred moderately often, while explicitly value-focused categories like *Malevolence* remained rare. Overall, the functionality-related and performance-oriented dimensions dominated both *Trust* and *Distrust* discourse, while ethical and normative aspects were less present.

Temporal Changes of Dimensions. Figure 3 shows how dimensions evolve in our data. Among dimensions of *Trust*, we observed *Competence* as the most dominant, maintaining the highest prevalence across the entire timeline. This suggests that perceptions of whether GenAI can effectively perform tasks were the central lens through which *Trust* was framed. *Familiarity* also remained consistently high, indicating that repeated exposure and growing experience with these systems reinforced *Trust* over time. *Reliability* appeared somewhat lower but still steady, pointing to its importance as a complementary indicator of technical soundness. *Integrity*, *Transparency*, and *Benevolence* remained marginal throughout. Their relative flatness across the timeline shows that while ethical values were present in discourse, they played a secondary role compared to functionality and user experience in social media discourse. The *Distrust* dimensions showed greater volatility and relatively lower presence in the data. *Unreliability* and *Incompetence* dominated this category, reflecting sustained concerns with system accuracy and task performance. *Opaqueness* and *Deception* also had modest rises over time, which can point to ongoing worries about transparency and hallucination. In contrast, *Malevolence* and *Dishonesty* remained minimal throughout, rarely gaining traction even in periods of heightened public debate.



(a) Dimensions of *Trust*

(b) Dimensions of *Distrust*

Figure 3: Normalized distribution of dimensions over time

Dimensions Co-occurence. To understand the relationships between dimensions, we calculated the *Jaccard index* for co-occurrence of different pairs of dimensions (Table F3 in Appendix). In the *Trust* posts, the strongest co-occurrence was between *Competence* and *Familiarity* (J = 0.76), followed by *Competence* with *Reliability* (0.66) and *Familiarity* with *Reliability* (0.52). Pairs such as *Competence-Transparency* and *Transparency-Familiarity* showed moderate overlap, suggesting that technical performance and clarity of system behavior were often discussed together. For *Distrust*, the dominant pair was *Incompetence-Unreliability* (J = 0.67), indicating that functionality failures were central to negative perceptions. Other notable co-occurrences were *Opaqueness-Unreliability*, and *Unreliability-Deception*, reflecting concerns about opacity and misinformation. Overall, Jaccard was higher for pairs of trust dimensions compared to distrust dimensions.

Classification Performance vs. Human Agreement. We noticed that the classification performance was lower than expected for several dimensions, including *Integrity, Transparency*, and *Unfamiliarity*, even with their best respective models. To further examine this, we analyzed how human annotators employed different dimensions during coding. We computed an aggregate ratio for each dimension reflecting both their frequency of use and the level of annotator agreement. Specifically, the ratio corresponds to the proportion of annotators assigning a given dimension to a post. For instance, if three out of five annotators label a post with *Competence*, the ratio for that dimension is 0.6 (3 out of 5), whereas dimensions not selected by any annotator receive a value of zero. These ratios were then aggregated across the annotated dataset for each dimension (Table F4). To assess classification performance across dimensions, we examined the relationship between aggregate annotation ratios and classification scores under the *majority* and *any* evaluation criteria. Our results (Figure F1 in Appendix) suggest a strong Pearson correlation between aggregate rates and classification performance (r = 0.83 in both approaches and statistically significant). While functionality-related dimensions such as *Competence, Reliability* and their semantic opposites demonstrate both a high agreement rate among human annotators and a high classification performance, ethical dimensions like *Integrity* and *Transparency* are suffering from both a low aggregated presence in human-annotators labels and a low classification performance.

4.3 Reasons of Turst & Trustors

Classification. Table 9 reports model performance on identifying the implied *Reason* for *Trust* or *Distrust*, as well as the *Trustor* in each post. Performance was relatively strong across models for both tasks. For *Reason*, GPT-5-mini performed best (F1 = 0.87), surpassing other models on both precision and recall. For *Trustor*, F1 scores were very close. GPT-5.1 and GPT-4.1 achieved the highest F1 (0.8), though GPT-5.1 showed slightly higher recall and precision. Given these results, we adopted GPT-5-mini for *Reason* and GPT-5.1 for *Trustor* to label the rest of our dataset with batch processing.

Reasons of Trust. Figure 4 shows the distribution of *Reasons* cited for posts labeled as *Trust*, *Distrust*, and *Both*. Across all three categories, 'personal experience using AI' was in the majority of posts, indicating that direct interaction with

Model	Precision	Recall	F1-Score
Reasons of Trust			
GPT-5.2	0.90	0.80	0.85
GPT-5.1	0.86	0.75	0.79
GPT-5	0.90	0.82	0.85
GPT-5-mini	0.91	0.85	**0.87**
GPT-5-nano	0.91	0.77	0.82
GPT-4o	0.91	0.78	0.83
GPT-4o-mini	0.87	0.71	0.78
GPT-4.1	0.89	0.85	0.86
GPT-4.1-mini	0.88	0.79	0.83
GPT-4.1-nano	0.85	0.15	0.22
Trustors			
GPT-5.2	0.80	0.79	0.78
GPT-5.1	0.81	0.82	**0.80**
GPT-5	0.80	0.79	0.78
GPT-5-mini	0.79	0.81	0.79
GPT-5-nano	0.78	0.78	0.78
GPT-4o	0.80	0.79	0.79
GPT-4o-mini	0.77	0.79	0.76
GPT-4.1	0.80	0.81	0.80
GPT-4.1-mini	0.78	0.81	0.78
GPT-4.1-nano	0.69	0.67	0.67

Table 9: Classification for Reason and Trustor

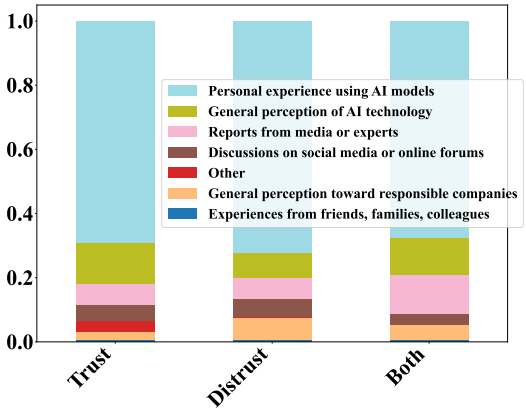


Figure 4: Distribution of *Reasons*

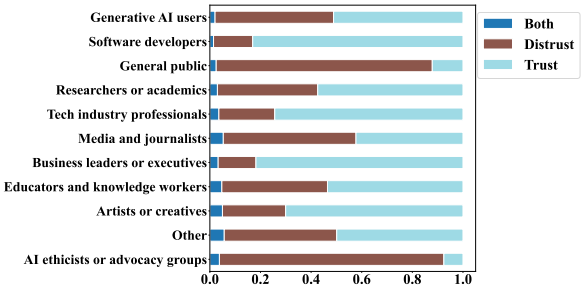


Figure 5: Distribution of different *Trustor* groups

GenAI was the primary driver of both *Trust* and *Distrust*. Among secondary reasons, 'general perception of AI' appeared more frequently in *Trust* posts, while 'general perception toward responsible companies' were more common in *Distrust*. 'Reports from media or experts' were cited as the secondary reason for posts tagged as *Both*. 'Discussions on social media,' and 'experience from friends, families, colleagues' were minimal, suggesting these external factors play only a supplementary role. Overall, these results reinforce that *Trust* and *Distrust* in GenAI are predominantly framed through firsthand use, with external narratives adding nuance but less weight (OpenAI, 2025).

Trustors. Figure 5 shows the breakdown of labels within each *Trustor* group. Among business leaders, academics, software developers, and tech professionals, *Trust* outweighs *Distrust*. In contrast, *Distrust* is more prevalent among AI ethicists, the general public, media & journalists (although smaller). Generative AI users (the largest group), and educators & knowledge workers were distinctive in showing relatively balanced levels of *Trust* and *Distrust*, suggesting a more cautious or divided perspective. These patterns show that while direct user experience drives the overall framing of (dis)trust, different stakeholders emphasize contrasting evaluations, highlighting the sociotechnical complexity of GenAI discourse.

5 Discussion

Trust and Distrust Are Not Static. Our analysis shows that *Trust* and *Distrust* co-exist in GenAI discourse, each accounting for a substantial share of posts, with local surges triggered by model releases. This dynamic mirrors survey findings that public attitudes toward AI remain mixed, balancing optimism with concern (Pew Research Center, 2025). It echoes work showing that breakthrough announcements coincide with intensified debate and media coverage (Ng and Chow, 2024; Leiter et al., 2024). Rather than converging toward stable positions, online discourse repeatedly oscillates between optimism and skepticism, showing the evolving nature of debates about the technology's value and risks. This is consistent with process-oriented models of human–AI trust in the literature (McGrath et al., 2025). Dimen-

sions analysis reveals that discourse is anchored in functionality: *Competence*, *Reliability*, and *Familiarity* dominate, while ethical dimensions such as *Transparency*, *Integrity*, and *Benevolence* are less salient. This imbalance suggests that users prioritize whether GenAI "works" over whether it is "good," reflecting prior findings that functionality and reliability form the baseline for technology adoption (Jiang et al., 2025; Gefen et al., 2024). Ethical concerns, though visible, remain secondary in everyday discussions, surfacing primarily in moments of controversy or policy debate (Crawford, 2021). Dimension co-occurrence analysis shows that functional dimensions frequently appear together. Jaccard values are higher for *Trust* pairs than for *Distrust* pairs, suggesting users often combine multiple positive aspects of trustworthiness, while expressing distrust more narrowly.

Dimensions: Classification vs. Salience. The relationship between classification performance and aggregate annotation ratios (which reflect both dimension prevalence and annotator agreement) indicates that model performance is closely tied to the clarity and consistency with which dimensions are expressed. Dimensions such as *Competence* and *Reliability*, which exhibit higher aggregate ratios, also achieve stronger classification performance. This suggests that models perform best when signals are both frequent and consistently interpreted by annotators. In contrast, dimensions such as *Integrity* and *Transparency* appear less frequently and show weaker classification results. These dimensions are more abstract and value-driven, making them harder for both annotators and models to identify reliably. The observed performance gap, therefore, reflects not only computational limitations but also differences in how users articulate *(Dis)Trust* online: concrete performance-related concerns tend to be expressed explicitly, whereas ethical critiques often require broader contextual interpretation (Binns et al., 2018). Prior work shows that annotation disagreement in subjective tasks frequently reflects genuine interpretive variation rather than noise, and that model performance is constrained by the level of human consensus in the training data (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019). Our findings suggest that lower performance on certain dimensions reflects variability in human interpretation as much as modeling limitations.

Trust and Distrust Depend on Who. Trust in GenAI is not uniform but shaped by the identity and perspective of the *Trustor* (Balayn et al., 2024). Our analysis shows that business leaders, academics, software developers, media, and tech professionals express more *Trust* than *Distrust*, consistent with survey evidence that experts tend to be more optimistic about AI than the general public (Gramlich, 2025). By contrast, AI ethicists and the general public lean more toward *Distrust*, reflecting critical perspectives grounded in ethical concerns and lived experience, echoing national survey findings of a widening AI optimism–skepticism divide (Schumann, 2025). Educators and knowledge workers show almost equal levels of *Trust* and *Distrust*, suggesting a more cautious or divided perspective. This aligns with prior work showing ambivalence in professional settings where AI intersects with labor and pedagogy (Buhmann and Fieseler, 2023). Results for *Reasons* highlight personal experience as the dominant basis for *Trust* or *Distrust*, outweighting mentions of external factors such as media reports and AI companies. This is consistent with usage studies by OpenAI and Anthropic (OpenAI, 2025; Anthropic, 2025), and previous research in HCI (Razi et al., 2024), which found most LLM interactions to be direct and experiential.

Governance, Design, and Literacy. We show that *Trust* and *Distrust* co-exist, and rather than converging toward a stable consensus, they continuously fluctuate. Governance frameworks and design practices must therefore treat both as independent forces, echoing the "trust and distrust paradox" where confidence and skepticism are expressed simultaneously (Ou and Sia, 2009). Our finding that functionality-oriented dimensions dominate over ethical ones raises questions about alignment with theoretical frameworks of *Trust*. Previous work (Mayer et al., 1995; Glikson and Woolley, 2020; Lee and See, 2004) incorporates both cognitive and affective factors that shape *Trust*, emphasizing ability, integrity, and benevolence as fundamental foundations of *Trust* that co-exist. Yet in public discourse, ability-based judgments dominate, while *Integrity* and *Benevolence* are far less salient. This suggests that classical definitions may not fully capture social media discourse on GenAI, how users articulate *Trust* in GenAI, or that users implicitly collapse ethical

and normative concerns into judgments of performance. This pattern likely reflects the experiential nature of social media discourse, where users evaluate technologies through direct interaction rather than institutional or ethical reasoning, overlooking normative, ethical concerns. Our results thus invite reconsideration of whether *Trust* in AI, as expressed in social media narratives, aligns with or diverges from established theoretical frameworks. From a digital literacy perspective, our findings highlight a gap between expert concerns and everyday user priorities: while policymakers emphasize fairness, bias, and accountability, most users focus on whether GenAI "works" in practice. Here, "working" often reflects perceived usefulness, whether outputs appear coherent, helpful, or time-saving, rather than evaluation of factual accuracy, bias, or completeness. A response that "looks acceptable" may still contain subtle errors or normative distortions that go unnoticed in routine use. Strengthening AI literacy, therefore, requires making normative risks visible in ways that connect to everyday use, ensuring governance frameworks address both practical utility and epistemic quality. These findings underscore the sociotechnical complexity of GenAI and the need for more comprehensive AI governance (Long and Magerko, 2020; Jobin et al., 2019).

6 Limitations

Our study has several limitations. First, in the classification of dimensions, performance varied across labels. Categories such as *Competence* and *Reliability* were classified with higher accuracy, but abstract dimensions like *Transparency* and *Integrity* performed poorly. We investigated the overall presence of dimensions in the sample set and found the link between the scarcity of these dimensions in the data and classification performance. We retained these dimensions but interpret them cautiously. Future work could improve reliability with expert input or hierarchical schemes that separate functional from normative dimensions. Second, our data is only in English and comes from Reddit, which, while large and temporally rich, skews younger and technologically engaged. Thus, our findings reflect specific online communities rather than the general population, and cross-platform studies are needed for broader validation. While LLM-assisted classification is promising, models still struggle with ambivalence (*Both*) and abstract normative categories, which are difficult to detect in informal posts.

7 Ethical Considerations

All data collection and annotation procedures were conducted under institutional IRB approval. Participants provided informed consent prior to beginning the study and were clearly informed of their rights, including the ability to withdraw at any time without penalty. Annotators recruited through Prolific were compensated at an average rate of 10 USD/hour, which was above minimum wage at the time of study and consistent with recommended fair-pay practices for crowdsourced research. We also included mechanisms to support participant well-being, such as a help button in the annotation interface and direct communication channels with the research team for resolving concerns. All posts were collected from publicly accessible subreddits, and we excluded deleted or removed content to respect author intent. To minimize risks of re-identification, posts were annotated in de-identified form, and no usernames or direct links were shared. We recognize that research on *Trust* and *Distrust* in GenAI carries broader ethical stakes. Our findings describe how public perceptions evolve in online communities, but should not be misused as definitive judgments about the reliability or trustworthiness of GenAI systems themselves. Instead, this work contributes to transparent monitoring of public discourse, with the goal of informing responsible AI design, governance, and literacy initiatives.

Acknowledgment

This research was supported by an award from Drexel University's Areas of Excellence & Opportunity. We thank Farnaz Ghashami and Kshitij Kayastha for their contributions in the early stages of this project. We also thank Trevor Canfor-Dumas and Jesus Romero for their help with the literature review and annotation of sample data. Finally, we thank the Prolific participants whose time and engagement made this study possible.

References

Maram Abdaljaleel, Muna Barakat, Mariam Alsanafi, Nesreen A Salim, Husam Abazid, Diana Malaeb, Ali Haider Mohammed, Bassam Abdul Rasool Hassan, Abdulrasool M Wayyes,

Sinan Subhi Farhan, et al. 2023. Factors influencing attitudes of university students towards chatgpt and its usage: a multi-national study validating the tame-chatgpt survey instrument.

AI@Meta. 2024. Llama 3 model card.

Giuseppe Amato, Malte Behrmann, Frédéric Bimbot, Baptiste Caramiaux, Fabrizio Falchi, Ander Garcia, Joost Geurts, Jaume Gibert, Guillaume Gravier, Hadmut Holken, et al. 2019. Ai in the media and creative industries. *arXiv preprint arXiv:1905.04175*.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. https://www.anthropic.com/news/claude-3-family. Accessed: 2026-05-13.

Anthropic. 2025. How people use claude for support, advice, and companionship.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI magazine*, 36(1):15–24.

Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in ai-enabled systems: An hci perspective. *International Journal of Human–Computer Interaction*, 40(5):1251–1266.

Agathe Balayn, Mireia Yurrita, Fanny Rancourt, Fabio Casati, and Ujwal Gadiraju. 2024. An empirical exploration of trust dynamics in llm supply chains. *arXiv preprint arXiv:2405.16310*.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'it's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 Chi conference on human factors in computing systems*, pages 1–14.

David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

Bernd Blöbaum et al. 2016. Trust and communication in a digitized world. *Models and Concepts of Trust Research. Heidelberg et al.: Springer*.

Diego De Siqueira Braga, Marco Niemann, Bernd Hellingrath, and Fernando Buarque De Lima Neto. 2018. Survey on computational trust and reputation models. *ACM Computing Surveys (CSUR)*, 51(5):1–40.

Alexander Buhmann and Christian Fieseler. 2023. Deep learning meets deep democracy: Deliberative governance and responsible innovation in artificial intelligence. *Business Ethics Quarterly*, 33(1):146–179.

José Siqueira de Cerqueira, Kai-Kristian Kemell, Rebekah Rousi, Nannan Xi, Juho Hamari, and Pekka Abrahamsson. 2025. Mapping trustworthiness in large language models: A bibliometric analysis bridging theory to practice. *arXiv preprint arXiv:2503.04785*.

Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Laleh Davoodi and Jozsef Mezei. 2024. A large language model and qualitative comparative analysis-based study of trust in e-commerce. *Applied Sciences*, 14(21):10069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Angelika Dimoka. 2010. What does the brain tell us about trust and distrust? evidence from a functional neuroimaging study. *Mis Quarterly*, pages 373–396.

Andreas Duenser and David M Douglas. 2023. Who to trust, how and why: Untangling

ai ethics principles, trustworthiness and trust. *arXiv preprint arXiv:2309.10318*.

Forhan Bin Emdad, Benhur Ravuri, Lateef Ayinde, and Mohammad Ishtiaque Rahman. 2023. "ChatGPT, a friend or foe for education?" analyzing the user's perspectives on the latest AI chatbot via reddit. *ArXiv*, abs/2311.06264.

Emilio Ferrara. 2024. Genai against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science*, 7(1):549–569.

David Gefen, Shahin Jabbari, Rezvaneh Shadi Rezapour, Aria Pessianzadeh, Kshitij Kayastha, and Hilde Van den Bulck. 2024. The evolving meaning of trust and risk in reddit discourse about chatgpt. In *Americas' Conference on Information Systems*.

David Gefen, Barry A Leffew, Arik Ragowsky, and Rene Riedl. 2025. The importance of distrust in trusting digital worker chatbots. *Communications of the ACM*, 68(4):42–49.

Michael Gerlich. 2024. Exploring motivators for trust in the dichotomy of human—ai trust dynamics. *Social Sciences*, 13(5):251.

Aashish Ghimire and John Edwards. 2024. Generative ai adoption in classroom in context of technology acceptance model (tam) and the innovation diffusion theory (idt). *arXiv preprint arXiv:2406.15360*.

Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660.

John Gramlich. 2025. Q&a: Why and how we compared the public's views of artificial intelligence with those of ai experts. Pew Research Center (Short Reads). Accessed: 2025-04-03.

Yuting Guo and Abeed Sarker. 2023. Socbert: A pretrained model for social media text. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 45–52.

Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1):89–100.

D Harrison McKnight and Norman L Chervany. 2001. Trust and distrust definitions: One bite at a time. In *Trust in cyber-societies: Integrating the human and artificial perspectives*, pages 27–54. Springer.

Sascha Hauke, Sebastian Biedermann, Max Mühlhäuser, and Dominik Heider. 2013. On the application of supervised machine learning to trustworthiness assessment. In *TrustCom*, pages 525–534.

Dan Heaton, Elena Nichele, Jeremie Clos, and Joel Fischer. 2024. "ChatGPT says no": agency, trust, and blame in twitter discourses after the launch of chatgpt. *AI and Ethics*, pages 1–23.

Arthur Heitmann. 2025. arctic_shift: Making reddit data accessible to researchers, moderators and everyone else. Accessed: 2025-07-15.

Robin Hill. 2023. Are you a doomer or a boomer? *Communications of the ACM*.

Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434.

Michael C Horowitz, Lauren Kahn, Julia Macdonald, and Jacquelyn Schneider. 2024. Adopting ai: how familiarity breeds both trust and contempt. *AI & society*, 39(4):1721–1735.

Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in ai. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 624–635.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile

Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Weizheng Jiang, Dongqin Li, and Chun Liu. 2025. Understanding dimensions of trust in ai through quantitative cognition: Implications for human-ai collaboration. *PloS one*, 20(7):e0326558.

Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of ai ethics guidelines. *Nature machine intelligence*, 1(9):389–399.

Krishnaveni Katta. 2025. Analyzing user perceptions of large language models (llms) on reddit: Sentiment and topic modeling of chatgpt and deepseek discussions. *arXiv preprint arXiv:2502.18513*.

Sage Kelly, Sherrie-Anne Kaye, Katherine M White, and Oscar Oviedo-Trespalacios. 2025. What factors predict user acceptance of chatgpt for mental and physical healthcare: an extended technology acceptance model framework. *AI & SOCIETY*, pages 1–19.

Sunnie SY Kim, Q Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. " i'm not sure, but...": Examining the impact of large language models' uncertainty expression on user reliance and trust. In *Proceedings of the 2024 ACM conference on fairness, accountability, and transparency*, pages 822–835.

Sunnie SY Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, ai, and context: Understanding end-users' trust in a real-world computer vision application. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 77–88.

Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on ai—an experimental study on the extent and costs of overreliance on ai. *Computers in Human Behavior*, 160:108352.

Nancy K Lankton and D Harrison McKnight. 2011. What does it mean to trust facebook? examining technology and interpersonal trust beliefs. *ACM SIGMIS Database: The DATABASE for Advances in Information Systems*, 42(2):32–54.

Nancy K Lankton, D Harrison McKnight, and John Tripp. 2015. Technology, humanness, and trust: Rethinking trust in technology. *Journal of the association for information systems*, 16(10):1.

John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. 2024. Chatgpt: A meta-analysis after 2.5 months. *Machine Learning with Applications*, 16:100541.

Roy J Lewicki, Daniel J McAllister, and Robert J Bies. 1998. Trust and distrust: New relationships and realities. *Academy of management Review*, 23(3):438–458.

Weixin Liang, Yaohui Zhang, Mihai Codreanu, Jiayu Wang, Hancheng Cao, and James Zou. 2025. The widespread adoption of large language model-assisted writing across society. *arXiv preprint arXiv:2502.09747*.

Magnus Liebherr, Mohamed Basel Almourad, Sameha AlShakshi, Christian Montag, Guandong Xu, Raian Ali, and Ala Yankouskaya. 2025. Developing and validating the attitudes toward large language models scale.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Duri Long and Brian Magerko. 2020. What is ai literacy? competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–16.

Niklas Luhmann. 2018. *Trust and power*. John Wiley & Sons.

Jorge López and Stephane Maag. 2015. Towards a generic trust management framework using a machine-learning-based trust model.

In *2015 IEEE Trustcom/BigDataSE/ISPA*, volume 1, pages 1343–1348.

Gjorgji Madjarov, Dragi Kocev, Dejan Gjorgjevikj, and Sašo Džeroski. 2012. An extensive experimental comparison of methods for multi-label learning. *Pattern recognition*, 45(9):3084–3104.

Andrew Marantz. 2024. Among the a.i. doomsayers. *The New Yorker*.

Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734.

Melanie J McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. 2025. Collaborative human-ai trust (chai-t): A process framework for active management of trust in human-ai collaboration. *Computers in Human Behavior: Artificial Humans*, page 100200.

D Harrison McKnight and Norman Chervany. 2001. While trust is cool and collected, distrust is fiery and frenzied: A model of distrust concepts.

D Harrison McKnight, Nancy K Lankton, Andreas Nicolaou, and Jean Price. 2017. Distinguishing the effects of b2b information quality, system quality, and service outcome quality on trust and distrust. *The Journal of Strategic Information Systems*, 26(2):118–141.

Harrison McKnight, Michelle Carter, and Paul Clay. 2009. Trust in technology: Development of a set of constructs and measures. *MIS Quarterly*.

Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A systematic review on fostering appropriate trust in human-ai interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing*, 1(4):1–45.

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Reuben Ng and Ting Yu Joanne Chow. 2024. Powerful tool or too powerful? early public discourse about chatgpt across 4 million tweets. *Plos one*, 19(3):e0296882.

OpenAI. 2024. Models overview: GPT-4, GPT-4o, and GPT-3.5. Accessed: 2025-09-27.

OpenAI. 2025. How people are using chatgpt.

Carol Xiaojuan Ou and Choon Ling Sia. 2009. To trust or to distrust, that is the question: Investigating the trust-distrust paradox. *Communications of the ACM*, 52(5):135–139.

Eva Paraschou, Maria Michali, Sofia Yfantidou, Stelios Karamanidis, Stefanos Rafail Kalogeros, and Athena Vakali. 2025. Ties of trust: a bowtie model to uncover trustor-trustee relationships in llms. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1715–1728.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Paul A Pavlou and David Gefen. 2004. Building effective online marketplaces with institution-based trust. *Information systems research*, 15(1):37–59.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. Potato: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

A. Pessianzadeh, N. Sultana, H. Van den Bulck, D. Gefen, S. Jabbari, and R. Rezapour. 2026. Dataset of trust and distrust in generative AI.

Jenny Pettersson, Elias Hult, Tim Eriksson, and Tosin Adewumi. 2024. Generative ai and teachers-for us or against us? a case study. In *14th Scandinavian Conference on Artificial Intelligence SCAI 2024*, page 37.

Pew Research Center. 2025. How the u.s. public and ai experts view artificial intelligence.

Afsaneh Razi, Layla Bouzoubaa, Aria Pessianzadeh, John S Seberger, and Rezvaneh Rezapour. 2024. Not a swiss army knife: Academics' perceptions of trade-offs around generative artificial intelligence use. *arXiv preprint arXiv:2405.00995*.

Denise M Rousseau, Sim B Sitkin, Ronald S Burt, and Colin Camerer. 1998. Not so different after all: A cross-discipline view of trust. *Academy of management review*, 23(3):393–404.

Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-ai decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Sebastian Schuetz, Le Kuai, Mary C Lacity, and Zach Steelman. 2025. A qualitative systematic review of trust in technology. *Journal of Information Technology*, 40(1):55–76.

Megan Schumann. 2025. Survey highlights an emerging divide over artificial intelligence in the u.s. Rutgers School of Communication & Information. Accessed: 2025-02-10.

Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. 2023. Chatgpt and other large language models are double-edged swords.

Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International journal of human-computer studies*, 146:102551.

Marita Skjuve, Asbjørn Følstad, and Petter Bae Brandtzaeg. 2023. The user experience of chatgpt: Findings from a questionnaire study of early users. New York, NY, USA. Association for Computing Machinery.

Sonia Sousa, José Cravino, and Paulo Martins. 2023. Challenges and trends in user trust discourse in ai popularity. *Multimodal Technologies and Interaction*, 7(2):13.

Nicole Miu Takagi. 2023. Banning of ChatGPT from educational spaces: A Reddit perspective. In *NLP4DH & IWCLUL*, pages 179–194, Tokyo, Japan. Association for Computational Linguistics.

Grigorios Tsoumakas and Ioannis Katakis. 2008. Multi-label classification: An overview. *Data Warehousing and Mining: Concepts, Methodologies, Tools, and Applications*, pages 64–74.

University of Edinburgh Impact. 2023. Openai corporate chaos reveals the war between ai 'doomers' and 'boomers'.

Ruyuan Wan, Jaehyung Kim, and Dongyeop Kang. 2023. Everyone's voice matters: Quantifying annotation disagreement using demographic information. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14523–14530.

Jingwen Wang, Xuyang Jing, Zheng Yan, Yulong Fu, Witold Pedrycz, and Laurence T. Yang. 2020. A survey on trust evaluation based on machine learning. *ACM Comput. Surv.*, 53(5).

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, et al. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

Jie Xu, Kim Le, Annika Deitermann, and Enid Montague. 2014. How different types of users

develop trust in technology: A qualitative analysis of the antecedents of active and passive user trust in a shared technology. *Applied ergonomics*, 45(6):1495–1503.

Hamdi Yahyaoui and Aisha Al-Mutairi. 2016. A feature-based trust sequence classification algorithm. *Information Sciences*, 328:455–484.

Sergio Zapata, Facundo Gallardo, Gustavo Sevilla, Estela Torres, and Raymundo Forradellas. 2023. Trust evaluation in virtual software development teams using bert-based language models. *Journal of Computer Science and Technology*, 23(1):e04–e04.

Yixuan Zhang, Joseph D Gaggiano, Nutchanon Yongsatianchot, Nurul M Suhaimi, Miso Kim, Yifan Sun, Jacqueline Griffin, and Andrea G Parker. 2023. What do we mean when we talk about trust in social media? a systematic review. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–22.

Yixuan Zhang, Yimeng Wang, Nutchanon Yongsatianchot, Joseph D Gaggiano, Nurul M Suhaimi, Anne Okrah, Miso Kim, Jacqueline Griffin, and Andrea G Parker. 2024. Profiling the dynamics of trust & distrust in social media: A survey study. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24.

A Data Collection

Table A1 shows the list of selected subreddits with counts of raw extracted posts and cleaned posts (before removing short posts). Here is the list of keywords related to generative AI: *Large Language Model**, *LLM*, *LLMs*, *LLM**, *GPT**, *ChatGPT**, *Google Bard*, *LLaMA**, *Claude Anthropic*, *AI apocalypse*, *Claude*, *OpenAI*, *Gemini*, *Language Models*, *PaLM*, *Pathways Language Model*, *Fine-tuned Language Net*, *FLAN*, *LaMDA*, *BLOOM*, *Language Model for Dialogue Application*, *BigScience Large Open-science Open-access Multilingual Language Model*, *Mistral AI*, *Mixtral*, *Replika*, *Deepseek*, *Gen-AI*, *GenAI*, *generative AI*, *Gen AI*, *generative artificial intelligence*.

B Dimensions of Trust and Distrust

Table B1 provides the full list of dimensions extracted from the literature. This list was further refined for our study as explained in § 3.2.1.

Subreddit	Creation Date	#Raw Post	#Filtered Post
r/AIJobs	24-Oct-17	135	9
r/Claude	6-Dec-13	188	116
r/OpenSourceAI	29-Jan-21	599	128
r/AITools	24-Sep-20	639	78
r/NLP	31-Dec-08	1,231	3
r/MistralAI	13-Jun-23	1,397	380
r/PromptDesign	10-Jun-22	1,617	411
r/HuggingFace	29-Aug-21	1,663	349
r/Anthropic	4-Feb-23	2,058	808
r/ControlProblem	29-Aug-15	3,434	203
r/LanguageTechnology	10-Mar-10	4,177	895
r/AGI	28-Jan-08	4,534	675
r/Automate	26-Jun-12	5,182	415
r/PromptEngineering	26-Feb-21	5,279	2,454
r/MLQuestions	4-Mar-14	8,362	1,097
r/deepseek	23-Nov-23	8,510	3,449
r/reinforcementlearning	2-Mar-12	8,593	387
r/GPT3	17-Jul-20	9,078	1,978
r/deeplearning	27-Nov-11	11,964	1,446
r/ChatGPTCoding	6-Dec-22	14,802	5,333
r/compsci	24-Mar-08	15,105	285
r/GitHub	22-Oct-10	16,100	378
r/Bard	6-Feb-09	19,141	6,905
r/learnmachinelearning	23-Feb-16	37,431	4,434
r/Futurology	12-Dec-11	41,684	893
r/artificial	13-Mar-08	50,374	3,370
r/replika	14-Mar-17	52,756	9,687
r/programming	28-Feb-06	68,796	513
r/LocalLLaMA	10-Mar-23	70,812	21,377
r/OpenAI	11-Dec-15	73,004	18,262
r/ArtificialInteligence	25-Feb-16	75,496	11,007
r/singularity	28-Jan-08	77,729	9,111
r/MachineLearning	29-Jul-09	82,513	5,689
r/technology	25-Jan-08	128,322	13
r/webdev	25-Jan-09	136,992	2,011
r/midjourney	13-Mar-22	147,076	979
r/StableDiffusion	23-Jul-22	162,138	3,012
r/Entrepreneur	21-Aug-08	193,635	2,581
r/ChatGPT	1-Dec-22	402,353	109,448

Table A1: List of subreddits used for data collection

C Annotation Procedure

C.1 Annotation Sample

We sampled 3K posts from our dataset for crowdsourced annotation. Rather than relying on purely random sampling, we used a targeted strategy to ensure sufficient representation of the key classes of interest: *Trust* and *Distrust*. Specifically, we first applied two open-source LLMs (Mixtral and Llama) to 5K posts and retained posts where both models agreed on one of the following labels: *Trust*, *Distrust*, or *Both*. From this subset, we drew a stratified random sample such that the final annotated dataset consisted of 40% *Trust*, 40% *Distrust*, and 20% *Neither*. This approach ensured that the annotation set contained a meaningful proportion of posts relevant to our research focus

Dimensions	Papers
Reliability	(Zhang et al., 2023; McKnight et al., 2009; Bach et al., 2024; Braga et al., 2018; de Cerqueira et al., 2025)
Integrity	(McKnight et al., 2017; Harrison McKnight and Chervany, 2001; McKnight and Chervany, 2001; Braga et al., 2018; de Cerqueira et al., 2025)
Benevolence	(Lankton and McKnight, 2011; Lankton et al., 2015; Zhang et al., 2024; Kim et al., 2024; de Cerqueira et al., 2025)
Competence	(McKnight et al., 2017; Harrison McKnight and Chervany, 2001; McKnight and Chervany, 2001; Braga et al., 2018; Xu et al., 2014)
Functionality	(Xu et al., 2014; Lankton and McKnight, 2011; Lankton et al., 2015)
Security	(Zhang et al., 2023; Bach et al., 2024)
Privacy	(Zhang et al., 2023)
Transparency	(Bach et al., 2024; Braga et al., 2018; Klingbeil et al., 2024; Shin, 2021; Duenser and Douglas, 2023; de Cerqueira et al., 2025)
Usability	(Zhang et al., 2023; Xu et al., 2014)
Helpfulness	(McKnight et al., 2009; Lankton and McKnight, 2011; Lankton et al., 2015)
Accuracy	(Bach et al., 2024; Schuetz et al., 2025; Salimzadeh et al., 2024)
Ethical use	(Zhang et al., 2023)
Predictability	(Harrison McKnight and Chervany, 2001; McKnight and Chervany, 2001; Braga et al., 2018)
Responsiveness	(Zhang et al., 2023)
Relatability	(Bach et al., 2024)
Explainability	(de Cerqueira et al., 2025)
Appearance	(Xu et al., 2014)
Consistency	(McKnight et al., 2009)
Decision-making	(Salimzadeh et al., 2024; Klingbeil et al., 2024)
Comprehension	(Salimzadeh et al., 2024)
Appraisal	(Salimzadeh et al., 2024)
Availability	(Salimzadeh et al., 2024)
Relevance	(Salimzadeh et al., 2024)
Interpretability	(Salimzadeh et al., 2024)
Reliance	(Klingbeil et al., 2024; Salimzadeh et al., 2024; Zhang et al., 2024)
Algorithmic bias	(Klingbeil et al., 2024)
Contradiction	(Klingbeil et al., 2024)
Deception	(Klingbeil et al., 2024)
Complexity	(Salimzadeh et al., 2024)
Difficulty	(Salimzadeh et al., 2024)
Uncertainty	(Salimzadeh et al., 2024)
Familiarity	(Bach et al., 2024)
Malevolence	(Zhang et al., 2024; McKnight et al., 2017)
Deceit	(McKnight et al., 2017)
Incompetence	(McKnight et al., 2017)

Table B1: Initial list of *Trust* and *Distrust* dimensions from the literature

while still preserving diversity in the data. As a result, the label distribution in the annotated sample differs from that of the full collected dataset.

C.2 Annotation Guideline and Questions

The annotators were given the following guidelines and questions:

```
Task Introduction
The purpose of this annotation task is to understand
 how people discuss trust and distrust toward
Generative AI (GenAI) and the companies responsible
for its development:
- Trust is defined as a belief in the reliability,
competence, or integrity of GenAI technologies /
responsible companies, leading to positive
expectations about their performance and ethical
behavior.
- Distrust is defined as skepticism or concern about
 the reliability, competence, or ethical
implications of GenAI / responsible companies,
leading to negative expectations or cautious
behavior toward these technologies or companies.
You will identify the presence or absence of trust
or distrust in each text, as well as the entities
who are trustors (those who express trust or
distrust) of specific GenAI systems/ companies.
Additionally, you will annotate the dimensions of
trust or distrust, which include aspects such as
reliability, transparency, integrity, and deception.
Instruction
Please note that you need to answer all the
necessary questions for each text to move forward to
 the next one, otherwise you will see an error that
keeps you from moving to the next text.
This instruction is accessible throughout the
annotation through the ''Help'' button on the upper-
left side of the screen. Please make sure to use it
at any time you need to review the definitions and
instructions.
Identify Trust or Distrust
For each text, determine whether it primarily
expresses:
Trust
Distrust
Both
Neither
If the label is Neither, move on to the next post by
 skipping the following questions. If the label is
Trust, Distrust, or Both, continue.
Label Trust Dimensions
Here are the definitions of trust dimensions. Choose
 one or more that are explicitly or implicitly
mentioned:
- Reliability (Consistent and accurate performance):
 Consistency and accuracy of the information
provided by LLMs.
- Transparency (Openness about how AI works): Being
open and clear about how data is collected,
processed, and how LLM works.
- Familiarity (Previous experience with AI):
Previous experience with AI or tools that makes LLMs
 more familiar.
- Integrity (Ethical use of AI): AI tool following a
 set of (ethical) principles that the user finds
acceptable.
- Competence (Ability to perform tasks effectively):
 Having the capability, functionality, or features
to do what users want
- Benevolence (Good intentions behind AI): Belief
that LLMs want to do good to the user and have good
intentions.
Label Distrust Dimensions
Here are the definitions of distrust dimensions.
Choose one or more that are explicitly or implicitly
 mentioned:
- Unreliability (Inconsistent, inaccurate results):
Consistency and accuracy of the information provided
 by LLMs.
- Opaqueness (Lack of openness or explainability):
Being open and clear about how data is collected,
processed, and how LLM works.
- Unfamiliarity (Uncertainty due to new or
unfamiliar tech): Previous experience with AI or
tools that makes LLMs more familiar.
- Dishonesty (Ethical concerns about lack of
integrity, bias, manipulation): AI tool following a
set of (ethical) principles that the user finds
acceptable.
- Incompetence (Fails to perform tasks effectively):
 Having the capability, functionality, or features
to do what users want
- Deception (Misleading or fabricated responses):
Belief that LLM is dishonest and potentially
provides false information, e.g., hallucination and
make up stuff.
- Malevolence (Harmful or dangerous AI use): Belief
that LLMs have the intention to harm the user, don't
 want to do good to the user.
Identify Reasons
What reasons for trust or distrust are mentioned in
the text? Identify the options from the list
provided in the questions.
Identify Trustor
Trustor:The entity expressing trust or distrust.
```

Available options are:
- Generative AI users
- Software developers
- Researchers or academics
- Tech industry professionals
- General public
- Media and journalists
- Business leaders or executives
- AI ethicists or advocacy groups
- Artists or creatives
- Educators and knowledge workers
- Other

Additional Points

This annotation task focuses on Generative AI and large language models (LLMs) such as ChatGPT, and companies responsible for their development (e.g., OpenAI). If trust is directed toward unrelated factors (e.g., hardware issues, people, or unrelated software), those should not be annotated. Focus only on trust or distrust related to LLMs, GenAI technologies, and responsible companies.

Annotation 1: Guideline.

Question 1:

Does the post express any level of trust or distrust in generative AI models or companies (e.g., ChatGPT, Llama, Open AI, or similar)?
- Trust
- Distrust
- Both Trust and Distrust
- Neither, No trust or distrust

Question 2:

Which dimensions of trust are discussed or implied in the post? (Select all that apply)

Reliability, Transparency, Familiarity, Integrity, Competence, Benevolence, None of the above.

Question 3:

Which dimensions of distrust are discussed or implied in the post? (Select all that apply)

Unreliability, Opaqueness, Unfamiliarity, Dishonesty, Incompetence, Deception, Malevolence, None of the above.

Question 4:

What is the main reason for trust or distrust expressed in the post? (Select one primary reason)
- Personal experience using AI models
- Reports from media or experts
- Discussions on social media or online forums
- General perception of AI technology
- General perception toward responsible companies
- Experiences from friends, families, colleagues, etc. who used it
- Other

Question 5:

Who is expressing trust or distrust? (Trustor)
- Generative AI users
- Software developers
- Researchers or academics
- Tech industry professionals
- General public
- Media and journalists
- Business leaders or executives
- AI ethicists or advocacy groups
- Artists or creatives
- Educators and knowledge workers
- Other

Annotation 2: Main Questions.

How old are you?
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65 or older
- Prefer not to say

Which of the following genders do you most identify with?
- Male
- Female
- Non-binary/ third gender
- Others
- Prefer not to answer

Which of the following best describes you? Select all that apply.
- Asian
- Black or African American
- Hispanic or Latino
- Native American or Alaska Native
- Native Hawaiian or Pacific Islander
- White
- Other

What is the highest level of education you have completed?
- Some high school, no diploma
- High school diploma or GED
- Some college, no degree
- Associate (2-year) degree
- Bachelor's (4-year) degree
- Master's degree
- Doctorate degree
- Prefer not to say

Which of the following best describes your employment status?
- Employed full-time
- Employed part-time
- Student
- Disabled
- Retired
- Unemployed
- Prefer not to say

Which of the following best describes your total annual income?
- Under 30,000
- 30,000 to 49,999
- 50,000 to 74,999
- 75,000 to 99,999
- 100,000 to 149,999
- 150,000 or more
- Prefer not to say

Which political party do you typically support?
- Democratic Party
- Republican Party
- Libertarian Party
- Green Party
- Independent
- Other
- Prefer not to say

What is your native language?
- English
- Spanish
- Arabic
- Chinese
- Hindi
- Other
- Prefer not to say

How would you describe your level of comfort and proficiency with technology?
- Beginner: I have limited experience with technology and often need assistance.
- Intermediate: I am comfortable using common technology and can troubleshoot basic issues.
- Advanced: I am very comfortable with technology and can handle complex software or troubleshoot issues independently.
- Expert: I have extensive knowledge of technology, including programming, system configuration, and troubleshooting complex issues
- Prefer not to say

How familiar are you with large language models (LLMs) like ChatGPT, Llama, or others?
- Not familiar: I have never heard of or used large language models.
- Somewhat familiar: I've heard of large language models and know a little about what they do.
- Moderately familiar: I understand the basic functions of large language models and have some experience with them.
- Very familiar: I have a strong understanding of large language models and regularly use or study them.
- Expert: I have extensive knowledge and experience with large language models, possibly in a professional or research capacity.
- Prefer not to say

How often do you use LLMsbased models or chatbots?
- Never
- Rarely
- Occasionally
- Frequently
- Always
- Prefer not to say

If you are familiar with large language models e.g.

```
ChatGPT, how would you describe your level of trust
in them?
- I fully trust them
- I somewhat trust them
- I'm not sure
- I somewhat distrust them
- I fully distrust them
- Prefer not to say
```

Annotation 3: Demographic Questions.

C.3 Annotation Interface and Procedure

The study interface included five components: (1) an introduction page describing the study purpose, (2) an instruction page with detailed guidelines and definitions, (3) a question page with five items per post, (4) a demographic page with 12 questions, and (5) an experience page for feedback. The session concluded with an end page that submitted responses and redirected participants to Prolific. A Help button allowed access to instructions at any point. We piloted the interface in-house (3 participants, 50 posts each) and ran two Potato tests (5 and 20 participants) before deploying the main study. Each round recruited 100 Prolific participants to annotate 50 posts each, yielding 1,000 posts per round with five annotations per post. Three attention checks were embedded, and only responses from participants who passed all were retained. To replace missing data, three supplemental rounds (17, 17, and 12 participants) were added. Eligibility required U.S.-based English speakers with a 95–100% approval rate and 100–10,000 prior submissions. Participants were paid above minimum wage, with the research team available for support throughout.

D Prompts Used for Classification Tasks

This section provides the prompts used in each classification task.

```
Task Introduction
You are an advanced AI model trained for text
classification. The purpose of this annotation task
is to classify how people discuss trust and distrust
 toward Generative AI dimensions of trust in text.
We define trust as a belief in the reliability,
competence, or integrity of GenAI technologies,
leading to positive expectations about their
performance and ethical behavior. On the other hand,
 we define distrust as skepticism or concern about
the reliability, competence, or ethical implications
 of GenAI, leading to negative expectations or
cautious behavior toward these technologies.
You will identify the presence or absence of trust
or distrust in each text, as well as the dimensions
of trust or distrust, which include aspects such as
reliability, transparency, integrity and deception.
Your task is to classify a given text:
Classify the text into a high-level category based
on showing trust or distrust toward GenAI.Follow the
 label definitions strictly.

Classify the text into one of the following
categories in terms of showing trust or distrust
toward Generative AI:
- Trust: The text expresses belief in the
reliability, competence, or integrity of GenAI
technologies, leading to positive expectations about
 their performance and ethical behavior.
- Distrust: The text expresses skepticism or concern
 about the reliability, competence, or ethical
implications of GenAI, leading to negative
expectations or cautious behavior toward these
technologies or companies.
- Both: The text indicates both trust or distrust in
 some way. Such a text shows tendencies that suggest
 both trust and distrust at the same time.
- Neither: The text is not expressing trust or
distrust towards GenAI, or is irrelevant to the
context of our study.
Output Format
[Label]
Strictly follow this format and don't add any other
text.
Read the following text and provide the best labels:
 {text}
```

Prompt 1: Trust vs. Distrust

```
Task Instruction
You are an advanced AI model trained for text
classification..
We define trust as a belief in the reliability,
competence, or integrity of GenAI technologies,
leading to positive expectations about their
performance and ethical behavior.
On the other hand, we define distrust as skepticism
or concern about the reliability, competence, or
ethical implications of GenAI, leading to negative
expectations or cautious behavior toward these
technologies or companies.
Identify the dimensions of trust/distrust toward
GenAI in text based on {task1_label}.
Conditional instructions
if task1_label = Trust:
    Choose one or more of the following dimensions:
Reliability, Transparency, Familiarity, Integrity,
Competence, Benevolence
elif task1_label = Distrust:
    Choose one or more of the following dimensions:
Unreliability, Opaqueness, Unfamiliarity, Dishonesty
, Incompetence, Deception, Malevolence
elif task1_label = Both:
    Choose one or more of the dimensions from both
lists
else:  (Neither)
    No further classification is required. Simply
output Neither.
Output format instructions
After labeling the text with dimensions, add a brief
 cue that explains why you chose each of the
dimensions.
Use the following format for your response:
Dimensions: [Dimension 1, Dimension 2, ...]
Cues: [Cue 1, Cue 2, ...]
Strictly follow this format and don't add any other
text.
Read the following text and provide the best
dimensions: {text}
```

Prompt 2: Dimensions of Trust vs. Distrust

```
Task Instruction
You are an assistant in detecting the trustor of
trust or distrust towards Generative AI or Large
Language Models mentioned in the text.
We define trust as a belief in the reliability,
competence, or integrity of GenAI technologies/
responsible companies, leading to positive
expectations about their performance and ethical
behavior. On the other hand, we define distrust as
skepticism or concern about the reliability,
competence, or ethical implications of GenAI/
responsible companies, leading to negative
expectations or cautious behavior toward these
technologies or companies.
Trustor: The entity expressing trust or distrust.
Trustors could be users of GenAI, software
developers, or others.
You will identify the trustor of trust or distrust
in each text from the following options:
```

```
- Generative AI users
- Software developers
- Researchers or academics
- Tech industry professionals
- General public
- Media and journalists
- Business leaders or executives
- AI ethicists or advocacy groups
- Artists or creatives
- Educators and knowledge workers
- Other
Output format:
[Label ID ONLY e.g. 1-7]
Do not add any extra text

Here is the text: {text}
```

Prompt 3: Detection of Trustor

```
Task Instruction
You are an assistant in detecting the reasons for
trust or distrust towards Generative AI or Large
Language Models mentioned in the text.
We define trust as a belief in the reliability,
competence, or integrity of GenAI technologies/
responsible companies, leading to positive
expectations about their performance and ethical
behavior. On the other hand, we define distrust as
skepticism or concern about the reliability,
competence, or ethical implications of GenAI/
responsible companies, leading to negative
expectations or cautious behavior toward these
technologies or companies.
You will identify the reason of trust or distrust in
 each text from the following options:
- Personal experience using AI models
- Reports from media or experts
- Discussions on social media or online forums
- General perception of AI technology
- General perception toward responsible companies
- Experiences from friends, families, colleagues,
etc.
- Other
Output format:
[Label ID ONLY e.g. 1-7]
Do not add any extra text.

Here is the text: {text}
```

Prompt 4: Detection of Reasons

E Demographic Summary of Annotators

Table E1 summarizes the demographics of annotators whose responses were retained.

F Additional Results and Performance Analysis

Tables F1 anf F2 show the full results of dimeniosn classification. Table F3 shows the top co-occurrent pairs of dimensions in *Trust* and *Distrust*. Table F4 shows the aggregate ratios of dimensions labeled by human annotators. Figure F1 shows how human agreement rates and classification performance for dimensions relate to each other.

G Error Analysis

G.1 Common Misclassification Types

Table G1 shows where misclassifications by LLM tend to happen using our test set. Most misclassifications occur where *Neither* is wrongly predicted instead of *Trust* or *Distrust*. While *Trust* is wrongly predicted as *Distrust* 8 times, the reverse has occurred only once, suggesting that the classifier is successful overall in not mistaking these two concepts with each other.

G.2 Linguistic Analysis of Errors

To deepen our understanding of what causes misclassifications by our models, we manually reviewed every instance in the test set where the ground-truth label and the label predicted by the best model (GPT-5.1) were different, and took notes of where and how errors happened. Table G2 presents the main types of errors we found, and two samples for each class. We found a few samples where the model's label seemed to be more appropriate than the humans'; however, these cases were relatively rare and typically involved ambiguous language or borderline instances that allowed for multiple plausible interpretations.

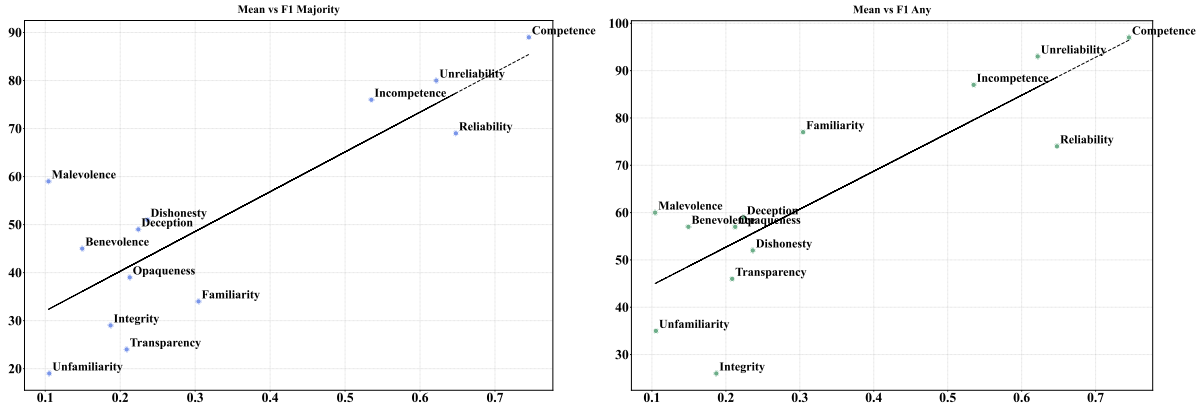
G.3 Zero-Shot vs Few-Shot

We noticed in Table 3 that the performance of zero-shot and few-shot models matched, where adding the examples to the few-shot prompt did not boost the classification performance. We made two comparisons in particular, one between GPT-5.1 zero-shot and few-shot (since it was the best zero-shot model), and the other between GPT-5.1 zero-shot and GPT-5-mini few-shot (the two best models in each category with the same results). Figure G1 shows how these two models aligned with the ground-truth label when they disagreed with each other. In comparison between GPT-5.1 zero-shot and few-shot, we observed that zero-shot is more aligned with the correct label when the label is *Trust* or *Distrust*. On the other hand, the few-shot is more accurate when the correct label is *Both* or *Neither*. This suggests that adding examples slightly improved the model's performance in understanding nuanced and mixed viewpoints, while zero-shot was capable of detecting more straightforward stances. When comparing GPT-5.1 zero-shot and GPT-5-mini few-shot (best model in each type), we observed that zero-shot was better in detecting *Trust* and *Neither*, while few-shot was overall better when the label was *Distrust* or *Both*. Overall, this pattern indicates a trade-off between precision in clear, polarized stances and sensitivity to ambiguity, with prompting strategy systematically shaping how models resolve this distinction.

	Question	#	(%)
Age	18-24	25	10.9%
	25-34	77	33.5%
	35-44	47	20.4%
	45-54	45	19.6%
	55-64	25	10.9%
	65 or older	10	4.3%
	Prefer not to say	1	0.4%
Gender	Male	84	36.5%
	Female	143	62.2%
	Non-binary / third gender	2	0.9%
	Other	0	0.0%
	Prefer not to say	1	0.4%
Race	Asian	4	1.7%
	Black or African American	64	27.8%
	Hispanic or Latino	10	4.3%
	Native American or Alaska Native	1	0.4%
	Native Hawaiian or Pacific Islander	1	0.4%
	White	146	63.5%
	Other	3	1.3%
	Prefer not to say	1	0.4%
Education	Some high school, no diploma	3	1.3%
	High school diploma or GED	14	6.1%
	Some college, no degree	26	11.3%
	Associate (2-year) degree	15	6.5%
	Bachelor's (4-year) degree	95	41.3%
	Master's degree	66	28.7%
	Doctorate degree	9	3.9%
	Prefer not to say	2	0.9%
Employment	Employed full-time	133	57.8%
	Employed part-time	59	25.7%
	Student	6	2.6%
	Disabled	2	0.9%
	Retired	6	2.6%
	Unemployed	22	9.6%
	Prefer not to say	2	0.9%
Income	$150,000 or more	22	9.6%
	$75,000 to $99,999	42	18.3%
	$100,000 to $149,999	49	21.3%
	$50,000 to $74,999	39	17.0%
	$30,000 to $49,999	33	14.3%
	Under $30,000	39	17.0%
	Prefer not to say	6	2.6%

	Question	#	(%)
Political Party	Democratic Party	92	40.0%
	Republican Party	87	37.8%
	Libertarian Party	4	1.7%
	Green Party	1	0.4%
	Independent	36	15.7%
	Other	3	1.3%
	Prefer not to say	7	3.0%
Native Language	English	229	99.6%
	Spanish	0	0%
	Chinese	0	0%
	Hindi	0	0%
	Arabic	0	0%
	Other	0	0%
	Prefer not to say	1	0.4%
Tech Level	Expert	38	16.5%
	Advanced	108	47.0%
	Intermediate	78	33.9%
	Beginner	5	2.2%
	Prefer not to say	1	0.4%
LLM Familiar	Not familiar	7	3.0%
	Somewhat familiar	31	13.5%
	Moderately familiar	83	36.1%
	Very familiar	90	39.1%
	Expert	18	7.8%
	Prefer not to say	1	0.4%
LLM Use	Never	9	3.9%
	Rarely	29	12.6%
	Occasionally	56	24.3%
	Frequently	106	46.1%
	Always	29	12.6%
	Prefer not to say	1	0.4%
Trust LLMs	Fully trust	27	11.7%
	Somewhat trust	134	58.3%
	Not sure	32	13.9%
	Somewhat distrust	32	13.9%
	Fully distrust	3	1.3%
	Prefer not to say	2	0.9%

Table E1: Summary of Annotators' Demographics



(a) Majority approach ground-truth

(b) Any approach ground-truth

Figure F1: Relationship between human agreement rates and classification performance for dimensions

Type	Model	Dimensions													Average
		Ben	Com	Dec	Dis	Fam	Inc	Int	Mal	Opa	Rel	Tra	Unf	Unr	Avg F1
Zero-Shot	gpt-oss-20b	0.35	0.84	0.48	0.48	0.25	0.74	0.28	0.52	0.35	0.59	0.18	0.09	0.77	0.46
	Mixtral-8x7B	0.24	0.87	0.35	0.36	0.32	0.48	0.19	**0.59**	0.30	0.54	0.18	0.10	0.77	0.41
	Llama-4	0.36	**0.89**	0.45	0.41	0.25	0.75	0.22	0.51	0.27	0.43	0.19	0.14	0.74	0.43
	gpt-5.2-2025-12-11	0.38	0.87	0.44	0.39	0.28	0.72	0.21	0.37	0.27	0.53	0.15	**0.19**	0.78	0.43
	gpt-5.1-2025-11-13	0.34	0.88	0.44	0.39	0.29	0.71	0.17	0.41	0.32	0.60	0.10	0.16	0.73	0.43
	gpt-5-2025-08-07	0.32	0.88	0.43	0.42	0.32	0.74	**0.29**	0.52	0.28	0.42	0.20	0.11	0.67	0.43
	gpt-5-mini-2025-08-07	0.36	**0.89**	0.42	0.45	0.31	**0.76**	0.25	0.55	0.34	0.68	0.20	0.11	0.79	0.47
	gpt-5-nano-2025-08-07	0.33	0.87	0.46	0.40	0.29	0.75	0.19	0.46	0.36	0.58	0.22	0.11	0.76	0.44
	gpt-4o-2024-11-20	0.35	**0.89**	0.45	0.39	0.30	0.73	0.20	0.43	**0.39**	0.46	0.14	0.15	0.77	0.43
	gpt-4o-mini-2024-07-18	0.37	0.87	0.41	0.48	0.27	0.73	0.21	0.49	0.34	0.55	0.16	0.11	0.79	0.44
	gpt-4.1-2025-04-14	0.41	0.88	0.48	0.42	0.31	0.75	0.26	0.43	0.37	0.60	0.14	0.12	0.78	0.46
	gpt-4.1-mini-2025-04-14	0.31	0.88	0.47	0.35	0.32	0.70	0.23	0.41	0.32	0.52	0.10	0.10	0.77	0.42
	gpt-4.1-nano-2025-04-14	0.31	0.81	0.36	0.29	0.31	0.48	0.18	0.54	0.25	0.59	0.20	0.07	0.75	0.39
Few-Shot	gpt-oss-20b	0.38	0.88	0.46	0.35	0.28	0.69	0.24	0.54	0.34	0.63	0.22	0.09	0.76	0.45
	Mixtral-8x7B	0.20	0.88	0.25	0.41	**0.34**	0.66	0.17	0.45	0.27	0.59	0.21	0.12	0.78	0.41
	Llama-4	0.37	**0.89**	0.42	0.38	0.22	0.72	0.22	0.43	0.29	0.48	0.19	0.04	0.72	0.41
	gpt-5.2-2025-12-11	0.41	0.87	0.45	0.35	0.27	0.74	0.20	0.37	0.32	0.53	0.16	0.11	0.78	0.43
	gpt-5.1-2025-11-13	0.38	0.88	0.47	0.36	0.32	0.71	0.19	0.41	0.32	**0.69**	**0.24**	0.17	0.77	0.45
	gpt-5-2025-08-07	0.29	**0.89**	0.46	0.38	**0.34**	0.75	0.27	0.54	0.31	0.50	0.22	0.15	0.74	0.45
	gpt-5-mini-2025-08-07	0.36	**0.89**	0.45	0.42	**0.34**	0.73	0.25	0.51	0.34	0.57	0.19	0.15	**0.80**	0.46
	gpt-5-nano-2025-08-07	0.29	0.86	0.42	0.27	0.32	0.70	0.21	0.53	0.32	0.62	0.19	0.09	0.77	0.43
	gpt-4o-2024-11-20	0.32	0.88	0.45	0.39	0.33	0.74	0.25	0.42	0.37	0.44	0.20	0.11	0.78	0.44
	gpt-4o-mini-2024-07-18	**0.45**	0.88	0.44	**0.51**	0.30	0.73	0.20	0.44	0.35	0.57	0.15	0.13	0.78	0.46
	gpt-4.1-2025-04-14	0.43	0.88	**0.49**	0.37	**0.34**	0.75	0.23	0.44	0.38	0.67	0.20	0.12	**0.80**	0.47
	gpt-4.1-mini-2025-04-14	0.30	0.86	0.45	0.34	0.32	0.67	0.23	0.35	0.35	0.61	0.17	**0.19**	0.75	0.43
	gpt-4.1-nano-2025-04-14	0.31	0.78	0.41	0.22	0.33	0.60	0.18	0.48	0.28	0.65	0.17	0.07	0.76	0.40

Table F1: Classification of dimensions of trust and distrust by zero-shot and few-shot models, and the scores used for our criteria under the Majority Criteria. Dimensions abbreviated as *Benevolence* (Ben), *Competence* (Com), *Deception* (Dec), *Dishonesty* (Dis), *Familiarity* (Fam), *Incompetence* (Inc), *Integrity* (Int), *Malevolence* (Mal), *Opaqueness* (Opa), *Reliability* (Rel), *Transparency* (Tra), *Unfamiliarity* (Unf), *Unreliability* (Unr).

Type	Model	Dimensions													Average
		Ben	Com	Dec	Dis	Fam	Inc	Int	Mal	Opa	Rel	Tra	Unf	Unr	Avg F1
Zero-Shot	gpt-oss-20b	0.35	0.90	0.49	0.40	0.43	0.84	0.12	0.33	0.36	0.59	0.11	0.18	0.81	0.46
	Mixtral-8x7B	0.47	0.95	0.35	0.45	0.70	0.48	0.12	0.33	0.53	0.56	0.25	0.15	**0.93**	0.48
	Llama-4	0.32	**0.97**	0.56	0.42	0.30	0.86	0.23	0.42	0.20	0.41	0.11	0.09	0.75	0.43
	gpt-5.2-2025-12-11	0.27	0.94	0.48	0.47	0.39	0.77	0.15	0.55	0.20	0.52	0.11	0.09	0.79	0.44
	gpt-5.1-2025-11-13	0.47	0.95	0.51	0.29	0.75	0.68	0.07	0.49	0.22	0.59	0.04	0.13	0.74	0.46
	gpt-5-2025-08-07	0.24	0.96	0.41	0.47	0.57	**0.87**	0.14	0.32	0.35	0.39	0.15	0.17	0.68	0.44
	gpt-5-mini-2025-08-07	0.41	**0.97**	0.53	**0.52**	0.73	**0.87**	0.20	0.36	0.52	0.68	0.23	0.23	0.84	0.55
	gpt-5-nano-2025-08-07	0.47	0.94	0.42	0.28	0.66	0.82	0.17	0.23	0.33	0.58	0.16	0.16	0.81	0.46
	gpt-4o-2024-11-20	0.43	**0.97**	**0.59**	0.45	0.75	0.77	0.10	0.53	0.44	0.44	0.13	0.20	0.82	0.51
	gpt-4o-mini-2024-07-18	0.29	0.94	0.40	0.38	0.48	0.83	0.09	0.42	0.50	0.53	0.11	0.29	0.91	0.48
	gpt-4.1-2025-04-14	0.29	0.96	0.52	0.36	**0.77**	0.81	0.10	0.49	0.48	0.58	0.09	0.30	0.85	0.51
	gpt-4.1-mini-2025-04-14	0.49	0.96	0.50	0.23	0.64	0.74	0.07	0.52	0.43	0.51	0.08	0.20	0.80	0.47
	gpt-4.1-nano-2025-04-14	0.42	0.86	0.28	0.18	0.60	0.41	0.07	0.36	0.48	0.60	0.26	**0.35**	0.81	0.44
Few-Shot	gpt-oss-20b	0.43	0.96	0.45	0.29	0.42	0.68	0.17	0.41	0.43	0.62	0.17	0.09	0.77	0.45
	Mixtral-8x7B	**0.57**	0.95	0.22	0.51	0.68	0.71	**0.26**	0.46	0.55	0.66	0.35	0.23	0.90	0.54
	Llama-4	0.39	**0.97**	0.54	0.41	0.22	0.76	0.22	0.46	0.19	0.44	0.21	0.10	0.73	0.43
	gpt-5.2-2025-12-11	0.37	0.93	0.44	0.41	0.41	0.73	0.17	0.53	0.25	0.51	0.18	0.07	0.77	0.44
	gpt-5.1-2025-11-13	0.44	0.96	0.48	0.32	0.67	0.67	0.13	0.50	0.26	0.72	0.12	0.14	0.75	0.48
	gpt-5-2025-08-07	0.26	**0.97**	0.40	0.41	0.58	0.78	0.16	0.42	0.45	0.47	0.21	0.18	0.74	0.46
	gpt-5-mini-2025-08-07	0.40	**0.97**	0.49	0.44	0.72	0.72	0.17	0.44	**0.57**	0.56	0.26	0.18	0.83	0.52
	gpt-5-nano-2025-08-07	0.50	0.92	0.38	0.16	0.64	0.68	0.21	0.32	0.47	0.61	0.24	0.23	0.79	0.47
	gpt-4o-2024-11-20	0.49	0.95	0.52	0.41	0.64	0.74	0.13	0.56	0.44	0.42	0.22	0.18	0.80	0.50
	gpt-4o-mini-2024-07-18	0.46	0.96	0.44	0.40	0.55	0.80	0.13	0.55	0.54	0.58	0.21	0.19	0.79	0.51
	gpt-4.1-2025-04-14	0.39	0.96	0.52	0.36	0.61	0.74	0.14	0.53	0.50	**0.74**	0.12	0.18	0.83	0.51
	gpt-4.1-mini-2025-04-14	0.54	0.92	0.47	0.24	0.62	0.61	0.10	**0.60**	0.38	0.63	0.20	0.16	0.76	0.48
	gpt-4.1-nano-2025-04-14	0.53	0.81	0.47	0.16	0.57	0.53	0.17	0.48	0.51	0.70	**0.46**	0.29	0.81	0.50

Table F2: Classification of dimensions of trust and distrust by zero-shot and few-shot models, and the scores used for our criteria under the Any Criteria. Dimensions abbreviated as *Benevolence* (Ben), *Competence* (Com), *Deception* (Dec), *Dishonesty* (Dis), *Familiarity* (Fam), *Incompetence* (Inc), *Integrity* (Int), *Malevolence* (Mal), *Opaqueness* (Opa), *Reliability* (Rel), *Transparency* (Tra), *Unfamiliarity* (Unf), *Unreliability* (Unr).

	dim_1	dim_2	co_count	Jaccard
Trust	Competence	Familiarity	55,275	**0.76**
	Competence	Reliability	47,191	0.66
	Familiarity	Reliability	36,061	0.52
	Competence	Transparency	34,206	0.47
	Familiarity	Transparency	27,765	0.43
	Transparency	Reliability	24,614	0.42
	Competence	Benevolence	20,546	0.29
	Familiarity	Benevolence	16,071	0.25
	Reliability	Benevolence	13,695	0.24
	Transparency	Benevolence	8,481	0.17
Distrust	Incompetence	Unreliability	39416	**0.67**
	Opaqueness	Unreliability	21518	0.36
	Deception	Unreliability	20967	0.36
	Unreliability	Unfamiliarity	20800	0.35
	Unreliability	Dishonesty	14744	0.25
	Incompetence	Deception	13374	0.26
	Incompetence	Unfamiliarity	13084	0.24
	Unfamiliarity	Opaqueness	12837	0.36
	Opaqueness	Incompetence	12327	0.23
	Dishonesty	Deception	10770	0.35

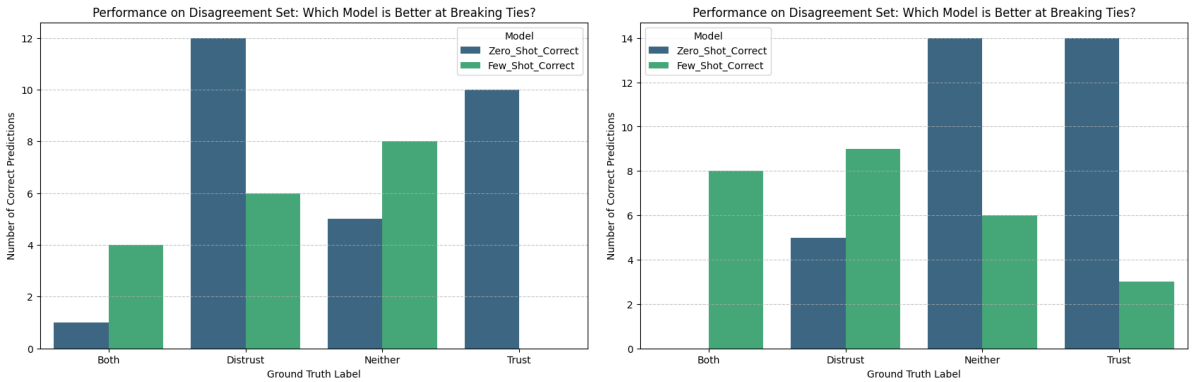
Table F3: Co-occurrent pairs of dimensions in *Trust* and *Distrust*

Dimension	Aggregate
Competence	0.75
Familiarity	0.30
Integrity	0.19
Reliability	0.65
Transparency	0.21
Benevolence	0.15
Deception	0.22
Dishonesty	0.24
Incompetence	0.54
Malevolence	0.10
Opaqueness	0.21
Unreliability	0.62
Unfamiliarity	0.11

Table F4: Aggregate ratios of dimensions tagged by human annotators

Actual (Ground Truth)	Predicted Label	Count
Trust	Neither	21
Neither	Trust	16
Distrust	Neither	14
Neither	Distrust	11
Trust	Distrust	8
Both	Distrust	7
Both	Trust	5
Trust	Both	2
Distrust	Trust	1

Table G1: Common mistakes with the best model



(a) zero vs few shot GPT-5.1

(b) zero shot GPT-5.1 vs few shot GPT-5-mini

Figure G1: Alignment of zero-shot and few-shot prompts with the ground-truth label in disagreement

Implicit vs. Explicit Trust Signals: Humans distinguish between using an LLM and trusting it, often labeling routine or exploratory use as *Neither*. In contrast, models frequently infer implicit *Trust* from usage, leading to over-classification of *Trust*.	***How can I use ChatGPT on Mac?*** *I'm a bit lost on what ChatGPT to download and what is suitable for my device? I need to use it on my laptop for work, and I want to help me write resumes and cover letters.*
	Transfer PDF into conversations. *I want ChatGPT to summarize papers. But most Journal papers I read are in a double-column format, which is hard to transfer to TXT. Are there any efficient ways for this on ChatGPT?*
Misinterpretation of Functional and Technical Queries: Posts framed as technical questions or support requests are typically labeled as *Neither* by humans. Models, however, often misclassify these by either treating them as *Trust* (due to engagement) or failing to detect dissatisfaction, reflecting difficulty separating instrumental use from attitudinal stance.	***Looking for recoms: What's the best PDF parser in ChatGPT?*** *I'm hoping to use ChatGPT's engine to process some PDFs, but I'm not sure which PDF parser to use. I've heard of some, but I was wondering if anyone can share experience or technical expertise. I appreciate any feedback or recommendations you might have!*
	Extremely low usage cap?? *I was working very briefly on tuning a GPT and I've apparently reached my cap within about 15 mins. Anyone else??*
Failure to Capture Implicit *Distrust* (Reliability and Frustration): Humans interpret cues such as frustration, failure ("not working"), or performance issues as signals of *Distrust*. Models frequently miss these implicit reliability concerns, instead treating them as neutral or purely transactional content.	***You've reached the current usage cap for GPT-4, try again after 2 hours.*** *Is there any way to get through this? Paying for what if I type only a few messages, I need to wait 2-3 hours? It was 0, then 1, now it's about 3 hours?*
	Still unable to get access to GPT Store. *This is ridiculous. I still do not have access to the GPT Store. There aren't any channels to report this issue and get it resolved. I just don't receive any updates that everyone else is getting.*
Difficulty with Nuance and Multi-Target Stances: Models struggle with complex or mixed expressions, including simultaneous *Trust* and criticism (*Both*), comparisons across different models or entities, and distinguishing *Trust* in LLMs vs. *Distrust* in companies or government. Humans capture these nuances more effectively, while models tend to collapse them into a single dominant label.	***Does the web browsing option actually work?*** *I have tried 3-4 different tasks for it to do, and it can't seem to do any of them. This last one seemed very simple, and it still produced no result. GPT4 is still worth the money, but the web version is useless. If anybody has any tips or has used it for something helpful, I'd be interested to hear your use cases.*
	ChatGPT 0, BingAI 0, Bard.Google 3!!! *I'm serious, THREE DAYS IN A ROW, my usual go-to ChatGPT couldn't answer me due to its cut-off date for data, Bing refused to answer me due to its perceived disrespectful term I was using (midgets), and in another case, simply couldn't access any data to answer me... YET BARD CAME THROUGH EVERY TIME with totally correct, reliable, and accurate information! Guess I have a new AI "companion" now!*
Edge Cases: Due to the nuanced nature of the posts, some disagreements occurred between human annotations and model predictions. For example, posts labeled as *Distrust* by annotators were sometimes classified as *Neither* by the model. These cases reflected borderline or ambiguous interpretations.	***LLM GPU forward compatibility [D]*** *So I heard an interview with a CoreWeave guy a while back saying that LLMs are not forward compatible with new GPU. Say it is designed to operate on A100's are not able to run efficiently on H100's, so A100's will be utilized for 5 to 10 years. Is this true?*
	ChatGPT Bios reliability *Is there a good reason that ChatGPT makes up simple biographies of relatively well-known people? Try it for yourself if you don't believe me.*

Table G2: Linguistic analysis of errors in classification