Social Intelligence Data Infrastructure: Structuring the Present and Navigating the Future

Minzhi Li 🦁⭐ **Weiyan Shi** 🌲 **Caleb Ziems** 🌲 **Diyi Yang** 🌲

🦁National University of Singapore

⭐Institute for Infocomm Research (I^2R), A*STAR

🌲Stanford University

li.minzhi@u.nus.edu weiyans@stanford.edu
cziems@stanford.edu diyiy@cs.stanford.edu

Abstract

As Natural Language Processing (NLP) systems become increasingly integrated into human social life, these technologies will need to increasingly rely on social intelligence. Although there are many valuable datasets that benchmark isolated dimensions of social intelligence, there does not yet exist any body of work to join these threads into a cohesive subfield in which researchers can quickly identify research gaps and future directions. Towards this goal, we build a *Social AI Data Infrastructure*[1], which consists of a comprehensive social AI taxonomy and a data library of 480 NLP datasets. Our infrastructure allows us to analyze existing dataset efforts, and also evaluate language models' performance in different social intelligence aspects. Our analyses demonstrate its utility in enabling a thorough understanding of current data landscape and providing a holistic perspective on potential directions for future dataset development. We show there is a need for multifaceted datasets, increased diversity in language and culture, more long-tailed social situations, and more interactive data in future social intelligence data efforts.

1 Introduction

> "*Data is a precious thing and will last longer than the systems themselves.*"
> — **Tim Berners-Lee**

As early as the 1920s, psychologists like Thorndike (1921) and Hunt (1928) considered social intelligence to be a distinct branch of intelligence that underlies all successful human interpersonal relationships. Many researchers now argue that social intelligence is a prerequisite of human-like Artificial Intelligence (Kihlstrom and Cantor, 2000; Erickson, 2009; Del Tredici et al., 2019; Radfar et al., 2020; Hovy and Yang, 2021; Williams et al., 2022). However, existing work still lacks a precise yet holistic definition for social intelligence in AI systems (Silvera et al., 2001).

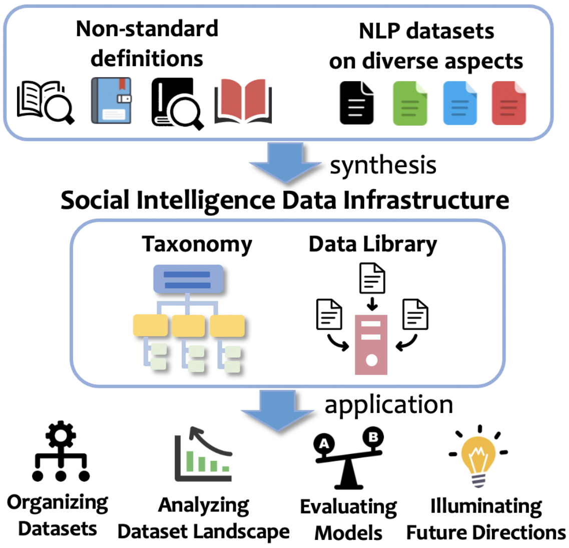
Non-standard definitions
NLP datasets on diverse aspects
synthesis
Social Intelligence Data Infrastructure
Taxonomy
Data Library
application
Organizing Datasets
Analyzing Dataset Landscape
Evaluating Models
Illuminating Future Directions

Figure 1: Our *Social Intelligence Data Infrastructure* gives a comprehensive overview and synthesis of social intelligence in NLP, with a theoretically grounded taxonomy and an NLP data library. Researchers can use our infrastructure to build and organize tasks, evaluate language models and derive future insights.

Prior studies define social intelligence by its cognitive aspect, or the ability to *understand* others (Barnes and Sternberg, 1989), while others emphasize the behavioral component by defining it as the ability to *interact* with other people (Hunt, 1928; Ford and Tisak, 1983). Both dimensions are relevant but incomplete, as social intelligence is multifaceted (Marlowe, 1986). Besides the narrow definitions, related empirical efforts, such as dataset collection and model development, also have isolated focuses and only address a single aspect of social intelligence (Fan et al., 2022). Therefore, there exists a pressing need for a holistic and synthesized definition for social intelligence to create an organized space for existing datasets. Without such organization, it is difficult to identify overar-

[1] Project Page

Findings of the Association for Computational Linguistics: ACL 2024, pages 2789–2805
August 11-16, 2024 ©2024 Association for Computational Linguistics

ching research questions and emerging trends for future exploration. Besides, although various social intelligence datasets have been proposed, the lack of data organization creates barriers for researchers to gain insights from previous work.

In light of this, we establish *Social Intelligence Data Infrastructure*, which consists of a comprehensive taxonomy for social intelligence and an organized data library of 480 NLP datasets (see Figure 1) to structure current data efforts and navigate future directions. The taxonomy (§2) formally defines various aspects of social intelligence, to introduce standardization and comprehensiveness to the definition of social intelligence in AI systems. The data library (§3.1) maps crawled datasets to different categories in our taxonomy, to provide structures for existing datasets. The taxonomy and data library can collectively aid researchers to identify existing dataset gaps and guide future dataset development for social intelligence.

Moreover, we demonstrate how the proposed *Social Intelligence Data Infrastructure* can be applied to gain insights into future dataset development with the following contributions:

- We perform distributional and temporal analysis (§3.3) to highlight overlooked categories and uncover emerging trends.
- We evaluate the zero-shot performance of Large Language Models (LLMs) (§4) on various social intelligence aspects defined in our taxonomy, to shed light on current models' capabilities and limitations.
- Finally, guided by the analysis and evaluation results, we discuss unfilled gaps and future directions for NLP dataset efforts on social intelligence (§5). We identify a need for multifaceted datasets, better diversity in language and culture, more long-tailed social situations, and more interactive data.

2 Social AI Taxonomy

To introduce a standardized and comprehensive definition of social intelligence, we propose *Social AI Taxonomy*, to capture diverse dimensions identified in previous work. As shown in Figure 2, different from previous categorization which is thematic with a focus on social understanding (Choi et al., 2023), our taxonomy considers the social interaction component and is hierarchical with three distinct types of social intelligence based on past literature: (1) cognitive intelligence, (2) situational intelligence, and (3) behavioral intelligence.

As identified in social cognitive theory (Bandura, 2009), these three intelligence types mutually influence each other to shape human behaviour. Prior work (Barnes and Sternberg, 1989; Kosmitzki and John, 1993) shows that these three intelligence types can comprehensively cover different factors and dimensions under social intelligence. Now we detail each intelligence type below.

2.1 Cognitive Intelligence

Definition. Cognitive intelligence refers to the use of verbal and nonverbal cues (Hunt, 1928) to understand others' mental states (Barnes and Sternberg, 1989). These include cold cognition about intents and beliefs, as well as hot cognition about emotions (Roiser and Sahakian, 2013), so our taxonomy decomposes cognitive intelligence into knowledge about ***intents***, ***beliefs***, and ***emotions***.

Significance. Cognitive intelligence includes the prerequisites for effective communication (Apperly, 2010) and many concrete NLP tasks (Langley et al., 2022). Task-oriented dialogue requires intent recognition (Jayarao and Srivastava, 2018; Wu et al., 2023), and mental health support demands an understanding of emotions (Peng, 2021; Singh and Srivastava, 2023). Broadly, Theory of Mind is a fundamental module in both both human (Premack and Woodruff, 1978) and artificial social intelligence (Rusch et al., 2020) that underlies downstream skills like stance awareness (see the left pillar of Figure 2).

2.2 Situational Intelligence

Definition. Situational intelligence refers to an awareness of the social context (Derks et al., 2007) and how this context informs the other pillars of cognition and behavior (*reciprocal interactions* in Figure 2). The literature tells us that social context includes: ***the social event*** itself (Sap et al., 2019b), as well as ***social and moral norms*** (Ziems et al., 2023a), ***culture*** and ***speaker information*** (Tigunova et al., 2021), all included in the *Taxonomy*.

Significance. Situational intelligence makes use of social context as the glue to bind the cognitive intelligence of mental states (§2.1) with a set of appropriate behaviors (§2.3), and serves as the foundation for decision making (Endsley, 1990). Given its centrality and its clear manifestations in decision making, situational intelligence has been a standard locus for social intelligence tests (Lievens

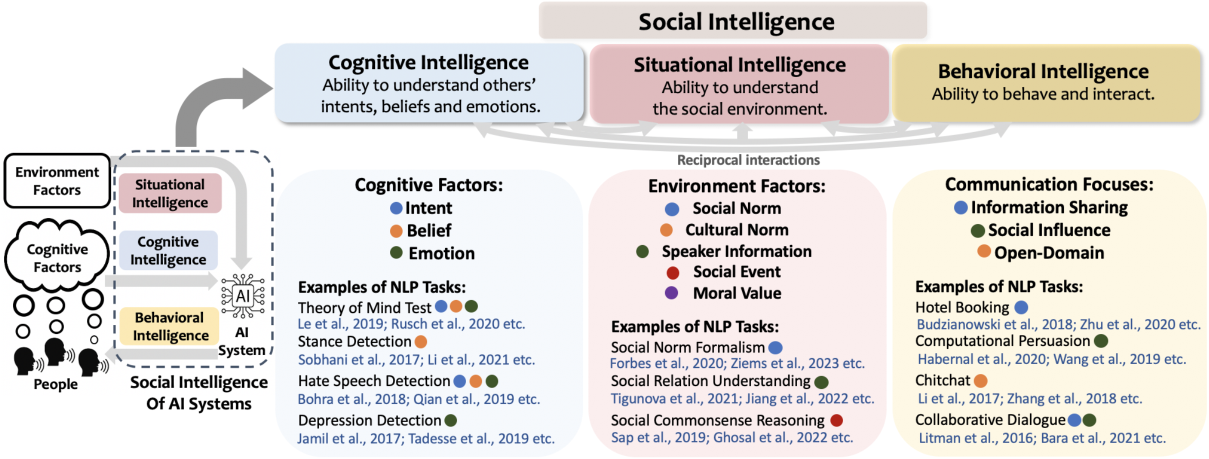
Social Intelligence
Cognitive Intelligence
Ability to understand others' intents, beliefs and emotions.
Situational Intelligence
Ability to understand the social environment.
Behavioral Intelligence
Ability to behave and interact.
Reciprocal interactions
Environment Factors
Cognitive Factors
People
Situational Intelligence
Cognitive Intelligence
Behavioral Intelligence
AI
AI System
Social Intelligence Of AI Systems
Cognitive Factors:
● Intent
● Belief
● Emotion
Examples of NLP Tasks:
Theory of Mind Test ●●●
Le et al., 2019; Rusch et al., 2020 etc.
Stance Detection ●
Sobhani et al., 2017; Li et al., 2021 etc.
Hate Speech Detection ●●●
Bohra et al., 2018; Qian et al., 2019 etc.
Depression Detection ●
Jamil et al., 2017; Tadesse et al., 2019 etc.
Environment Factors:
● Social Norm
● Cultural Norm
● Speaker Information
● Social Event
● Moral Value
Examples of NLP Tasks:
Social Norm Formalism ●
Forbes et al., 2020; Ziems et al., 2023 etc.
Social Relation Understanding ●
Tigunova et al., 2021; Jiang et al., 2022 etc.
Social Commonsense Reasoning ●
Sap et al., 2019; Ghosal et al., 2022 etc.
Communication Focuses:
● Information Sharing
● Social Influence
● Open-Domain
Examples of NLP Tasks:
Hotel Booking ●
Budzianowski et al., 2018; Zhu et al., 2020 etc.
Computational Persuasion ●
Habernal et al., 2020; Wang et al., 2019 etc.
Chitchat ●
Li et al., 2017; Zhang et al., 2018 etc.
Collaborative Dialogue ●●
Litman et al., 2016; Bara et al., 2021 etc.

Figure 2: Social AI taxonomy with three pillars: cognitive, situational and behavioral intelligence. We illustrate their respective roles in social interactions (left), and visualize their definitions and example NLP tasks (right).

and Chan, 2017; Hunt, 1928). The social context spans not only interpersonal factors like tie strength (Sap et al., 2019a; Cristani et al., 2011), but also cultural differences like those between high and low-context cultures (Hall and Hall, 1987; DeVito, 2016), and navigating these differences is essential for cross-domain and cross-cultural communication (Wawra, 2013). Studies have shown that incorporating these factors can lead to significant performance improvements in NLP systems (Rahimi et al., 2018; Wu et al., 2021).

2.3 Behavioral Intelligence

Definition. Behavioral intelligence refers to skills of successfully communicating and acting in a manner to attain social goals (Ford and Tisak, 1983) through (1) **information sharing** (Zhang et al., 2020), (2) **social influence** (Turner, 1991; Cialdini and Goldstein, 2004; Chawla et al., 2023; Weinstein, 1969), or (3) maintaining interpersonal relationships (Vernon, 1933; Moss and Hunt, 1927) via **open-domain** conversations (Huang et al., 2020). Our *Taxonomy* is organized around these three foci.

Significance. Behavioral intelligence has direct ramifications for human-human and human-AI interactions (see the right pillar of Figure 2). Task-oriented dialogue systems (Zhang et al., 2020) and collaborative AI partners (Bara et al., 2021) depend on successful information sharing, while other applications require engaging and personalized open-domain chit-chat (Zhang et al., 2018a). The capacity for social influence becomes relevant in human-AI teams (Bansal et al., 2021), where such skills can make use of advances in explainable AI systems (Angelov et al., 2021). Across these diverse applications, systems need to be equipped with social-behavioral skills like empathy (Rashkin et al., 2018), persuasion (Hunter et al., 2019), and transparency to build trust (Liao and Sundar, 2022).

2.4 Challenges in Measuring Intelligence

The three pillars of social intelligence are *not* mutually exclusive, nor are they readily isolated in social life, since they coordinate through reciprocal interactions (Figure 2). A situationally intelligent agent can better express cognitive intelligence, using cues from the social context to infer the mental states of others. The converse is also true that, by considering others' mental states, one can understand her role in shaping the social situation. Both intelligence can facilitate effective social actions.

Because of its dynamic nature, social intelligence may be outside the scope of what AI engineers can benchmark with any single static dataset. This is especially true as *situational* and *behavioral* pillars themselves are not static, but refer rather to an agent's ability to adapt into a social equilibrium. In our analysis and discussion, we will consider the degree to which benchmarks can reflect this dynamic nature, and whether existing datasets measure more than one type of social intelligence. For example, the COBRACORPUS (Zhou et al., 2023b) requires cognitive and situational intelligence to reason about offensive intents in different social contexts. Finally, we make suggestions for the design of data resources and the future of social AI.

3 Current Social NLP Data Landscape

With what granularity can existing data resources help researchers train and evaluate the core pillars of social intelligence in AI systems? How holistic is the landscape, and how sufficiently integrated are these pillars in the literature? To answer these questions, we leverage the *Social AI Taxonomy* to categorize existing NLP publications into a library of relevant datasets.

3.1 Data Library Construction

We use ACL Anthology data[2] crawled by Rohatgi et al. and Held et al. (2023), and set our time scope from year 2001 January to 2023 October as there are few datasets before 2001. We automatically collect social intelligence dataset papers by filtering titles and abstracts with keywords related to both (a) social intelligence and (b) dataset development (see Appendix D). The smaller size of this filtered pool allows us to manually curate papers, removing any surveys or irrelevant works on model development, annotation schemes, or annotation tools, which results in a curated set of 480 papers. For these papers, we scraped useful metadata, like title, url and publication year. We discuss the technically infeasibility of an exhaustive library in §6.

3.2 Metadata Annotation

We map the papers in our data library to the *Social AI Taxonomy*. Two authors reviewed the content in the paper and annotated them with type of intelligence and §2 subcategory (Cohen κ = 0.86 *cognitive*; 0.80 *situational*; 0.87 *behavioral*). The annotation is based on the main focus of the dataset. For example, if a work collects interactive dialogues solely for intent recognition purpose, we will classify it as cognitive intelligence instead of behavioral intelligence. Constructing the data library [3] illustrates how our theoretical taxonomy can be practically useful to organize datasets focusing on different aspects of social intelligence. On top of that, we also annotate other important attributes for each dataset by reviewing the paper contents: we annotate ***NLP Task***, ***Data Source*** (where the data was collected from), ***Annotation Strategy*** (how the labels were obtained), ***Generation Method*** (if the text comes from human, AI or both), ***Data Format*** (e.g. tweet, news article, dialogue etc.), ***Language***, ***Modality***, and ***Public Availability*** of the data.

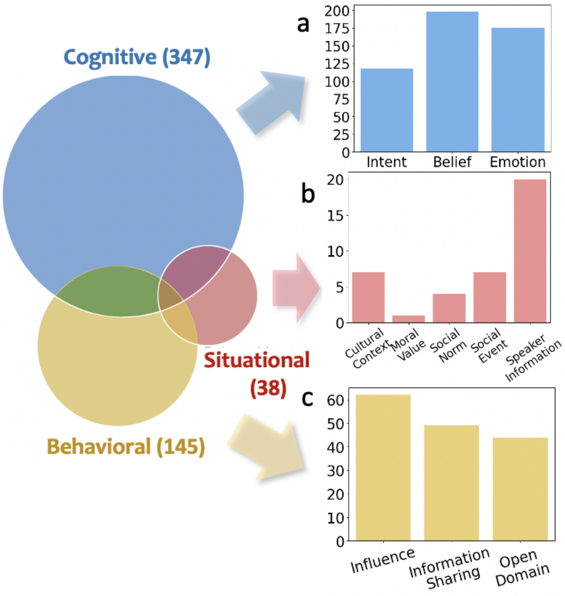

Figure 3: Distribution of three intelligence types (left) and frequency of different subcategories within cognitive, situational and behavioral intelligence (right).

3.3 Social NLP Data Landscape Highlights

By visualizing the distribution and temporal trend of the datasets in the data library, we obtain insights about the past and current NLP paradigm for dataset development on social intelligence. We discuss key results in this section and put more detailed analyses in Appendix A.

Topic Distribution Figure 3 shows most of the social NLP datasets focus on the cognitive aspect of social intelligence (64.2%), followed by behavioral aspect (22.7%), and least of all, its situational aspect (3.8%). Only a ***small set*** of datasets (9.4%) span ***multiple intelligence types***. We also visualize a detailed breakdown of different factors within each type. For cognitive and behavioral intelligence, papers are balanced across the respective subcategories. For situational intelligence, most datasets measure knowledge of speakers involved in the dialogue such as their demographics and social relations, and there are ***very few datasets on moral values and social norms***.

Temporal Topic Shift From Figure 4, we can better understand the temporal variation and shift of focus over time for each type of intelligence. We can see the ***onset of study on situational intelligence is later*** (2008) than the other two types (2001). For all three intelligence, the task of focus has become more ***specific and nuanced*** over the years. For example, early work on cognitive intelligence focused on general dialogue act classification

[2] https://aclanthology.org/
[3] Data library

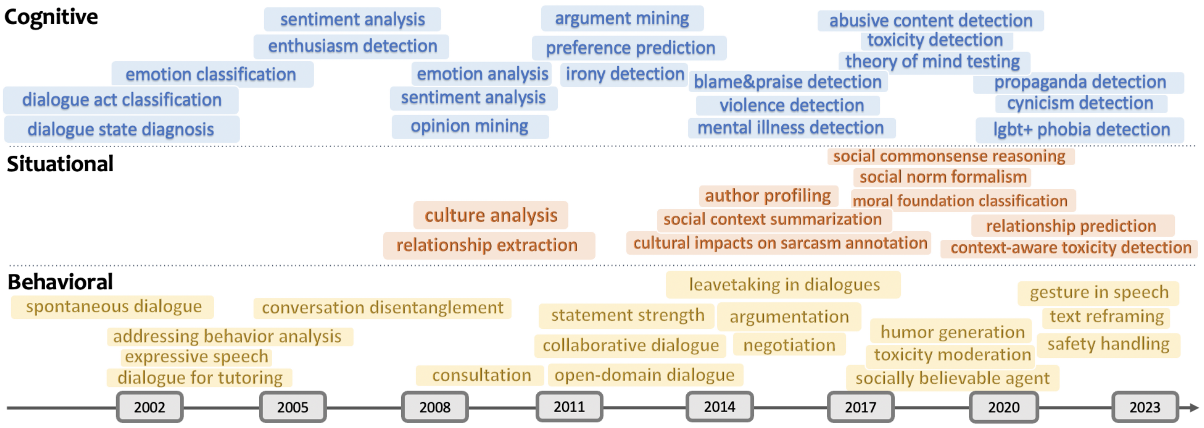

Figure 4: NLP tasks related to social intelligence over time. We show newly emerged topics based on the NLP Task field in our constructed data library for every three years. This is a non-exhaustive visualization (if number of distinct new topics for the period is more than three, we cap at three).

but recent studies are about more nuanced and challenging intents beyond literal meaning like sarcasm and irony understanding (Alnajjar and Hämäläinen, 2021; Frenda et al., 2023). Literature on behavioral intelligence began with tasks to identify effective or powerful written communication (Tan and Lee, 2014), and moved to more specific tasks like high-quality persuasive arguments for particular forms of negotiation (Ng et al., 2020; Chawla et al., 2021).

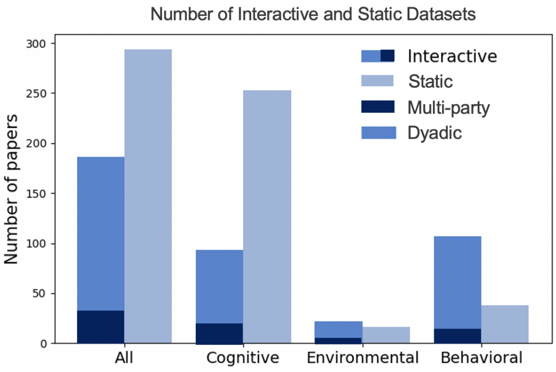

Figure 5: Number of papers with interactive or static data. We also visualize a breakdown of interactive data into dyadic and multi-party interactions.

Interactive vs Static Data We classify the data formats of surveyed datasets into two broad categories – interactive and static. Interactive data are those with information exchange like social media threads and daily conversations. Static data include (1) self-contained and topically focused texts written for a general audience (e.g. news and books) and (2) those that are part of information exchange but have no prior or subsequent context (e.g. a single post on Twitter or an utterance in a conversation). The difference in data format can affect the ability for language model to acquire social intelligence (Sap et al., 2022). Figure 5 shows that ***for cognitive aspects*** in particular, but not for other pillars, there are significantly ***more static datasets*** than interactive ones, which quantitatively confirms the trend observed by Sap et al. (2022). Moreover, within the interactive datasets, the ***proportion of multi-party modeling is small*** (18.4%).

Use of AI There are ***increasing number of works adopting AI*** for generating and annotating datasets related to social intelligence (before 2015: 3; after 2015: 32). We find that ***degree of adoption of AI for generation is higher*** than annotation. In recent work, researchers outsource generation completely by generating contents purely using AI (Zhou et al., 2023b) or simulating conversations between AI and AI (Lee et al., 2022). On the other hand, use of AI in annotating social intelligence data still remains in a hybrid stage (Jo et al., 2020) and AI usually plays a part in annotating simpler high-level components like themes (Maës et al., 2023).

4 Model Performance

Now we use our *Social AI Data Infrastructure* to evaluate current LLMs' performance on social intelligence and gain insights on models' strengths and limitations, shedding light on aspects which future social intelligence datasets need to address.

4.1 Experimental Setup

Dataset selection. For each of our taxonomic categories, we select two representative datasets, one simple and the other more challenging. Challenges

Intelligence	**Category**	**Dataset** (year)	**Task**	**Human**		**LLM**		
				Average	*Best*	*Claude*	*GPT − 4*	*Llama2*
Cognitive	Intent	SNIPS (2018)	query intent classification	82.5	**97.5**	78.8	95.0*	76.7
		iSarcasm (2018)	intended sarcasm detection	64.8	**90.9**	56.4	67.3*	55.5
	Belief	SemEvalT6 (2016)	stance detection on abortion	54.4	**84.2**	59.2*	76.7*	45.8
		WTWT (2020)	stance detection on merger and acquisition	61.4	**81.7**	47.5	55.8	75.0*
	Emotion	SemEvalT1 (2018)	emotion classification (11 classes)	78.9	**83.2**	78.5	80.9*	78.6
		GoEmotions (2020)	emotion classification (28 classes)	90.9	**93.1**	92.0*	90.2	92.6*
Situational	Social Situation	SocialIQa (2019)	commonsense reasoning with description	55.8	**80.0**	70.8*	70.8*	63.3*
		CICERO (2022)	commonsense reasoning with dialogues	77.2	85.3	79.3*	**86.7***	76.2
	Social Norm	NormBank (2023)	situational judgment	52.2	**71.7**	52.5*	60.8*	40.0
		MoralExceptQA (2022)	situational exception judgment	44.8	**93.2**	47.3*	50.0*	33.1
Behavioral	ChitChat	DailyDialogue (2017)	daily conversations†	50.0	-	74.7*	84.4*	**88.6***
		PersonaChat (2018)	chats conditioned on personas †	50.0	-	64.0*	**78.5***	64.0*
	Persuasion	Convincing Arguments (2016)	argument generation †	50.0	-	93.6*	**97.8***	93.1*
		PersuasionforGood (2019)	persuasive dialogue response generation †	50.0	-	74.3*	**96.3***	93.9*
	Therapy	Positive Reframing (2022)	reframing text in a positive way †	50.0	-	83.5*	**96.2***	92.0*
		Counsel-Chat (2020)	provide counseling to problems †	50.0	-	62.6*	**93.5***	84.6*
Multiple	Intent+Social Situation	CobraCorpus (2023)	contextual offensive statement detection	73.6	**95.0**	70.0	89.2*	50.0
	Intent+ Cultural Norm	CulturalNLI (2023)	culturally aware natural language inference	52.1	**72.3**	33.3	65.0*	41.7

Table 1: Human and LLMs' performance on classification (F1 scores) and generation tasks (% preferred over average human). Within each category, we select one simpler (top) and one more nuanced (bottom) dataset for comparison. LLM performance that exceeds average human performance is marked with * and best performance is **bolded**. † Generation tasks have average human performance of 50% by definition; best performance is not defined.

arise from factors like nuancedness (e.g., sarcastic intents), task granularity (e.g., emotion detection with more fine-grained classes), data scarcity (e.g., stance detection in the economic domain), long-tail data distributions (e.g., moral exception), and challenging data formats (e.g., persuasion in a dialogue setting) (see Appendix E). We select open-sourced and widely-used datasets with high citations for each category. Testing the model on the entire dataset can be computationally expensive and time-consuming so we adopt class-stratified sampling of 100 to 150 instances from the original test set (if available). We evaluate LLMs' zero-shot performance on sampled instances, and follow the recommended practice of prompting as described by Ziems et al. (2023c).

Metrics. For classification tasks, we choose F1 as the metrics. For generation tasks, we present both the original human response and the LLM response to human annotators, and calculate the preference percentage. As such, we can better understand current social capability of language models, in both absolute terms against the ground truth and relative terms compared to human.

Human performance. For classification, we report the best and average F1 scores for responses from three Amazon Mechanical Turk workers per test instance. For generation tasks in which we compute preferences, we define the average human performance as 50%, without defining the best human performance.

4.2 Result Analysis

Main Results. From Table 1, we can see that ***LLM performs better on simpler datasets than more nuanced ones***. For example, compared to straightforward query intent recognition (95.0 F1), the best performing LLM (GPT-4) struggles more with identifying the intended sarcasm (67.3 F1) when people convey an opposite meaning from what they literally said. Moreover, uncommon tasks with fewer datasets are more challenging, such as stance detection in the economic domain (most stance detection data is for political domain (Küçük and Can, 2020)), moral exceptions, language inference under different cultures, and so on. With more fine-grained definitions on labels, LLMs have better performance in classification as seen from a higher F1 on GoEmotions than SemEvalT1 with more emotion classes defined. Additionally, more social context in the data can also result in better performance: for instance, they achieve a higher F1 on the CICERO dataset with both social situation description and dialogue data, than the SocialIQa dataset with only a simple description.

Performance Comparison with Humans. Table 1 shows that, for every task there exists at least one LLM surpassing the average human performance. However, ***LLMs perform worse than best human performance on most tasks on cognitive and situational intelligence***. The gap between LLMs (e.g. GPT-4) and the best human performance is higher for more nuanced tasks (iSarcasm: 23.6 vs. SNIPS: 2.5), task in scarce domains (WTWT: 25.9 vs. SemEvalT6: 7.5) and more long-tailed situations (MoralExceptQA: 43.2 vs. NormBank: 10.9). On the other hand, ***LLMs exceed average human performance on behavioral intelligence tasks*** with percentage preferred more than 50% on all tasks. However, percentage preferred for LLMs (e.g. Claude) is lower in more dynamic and interactive situations (e.g. applying persuasion in dialogue: 74.3 vs. writing persuasive arguments: 93.6) with more constraints (e.g. with persona constraints: 64.0 vs. without persona constraints: 74.7) given. More qualitative analysis about human and LLMs performance is provided in Appendix C.

Multiple intelligence. LLMs in real-life social applications usually require multiple intelligence (e.g. interpreting intents under different cultural backgrounds) but they are still lacking in performance (CulturalNLI: 65.0). Table 1 shows they perform well for individual modules, so systems can utilize LLMs for individual modules which LLMs do exceptionally well in and combine them organically to build a strong holistic system (e.g. combine emotion recognition and positive reframing components for a counseling system).

5 Recommendations for the Future

From analyzing current data landscape (§3.3) and evaluating LLMs' performance (§4), we unveil the most challenging aspects of social intelligence that remain unaddressed by existing data resources or model capabilities. Guided by insights from our results, we discuss possible future directions for dataset development below.

5.1 Recommendations for Data Content

Future datasets should focus more on ***specific, nuanced, and long-tailed social situations***. The LLMs we evaluated fell short on nuanced tasks like sarcasm and moral exceptions. As human expressions are diverse and subtle, and vary with complex linguistic contexts (Cruse, 2004), it is crucial to model the *contextual complexity* to address the challenge of *ambiguity* in language. Long, multi-party interactions that go beyond conventional dialog or small groups are also critical for developing more sophisticated social intelligent systems that account for *variation in discourse structure* and *potential conflicts between individual and group objectives,* which produce novel social equilibria.

As shown in Figure 3, most datasets focus on just a single intelligence type. However, different intelligence types are not isolated but rather *reciprocal* in real-world applications (Fan et al., 2022). Thus, there should be ***more multifaceted datasets encompassing multiple intelligence types*** to promote holistic benchmarking.

Moreover, there is a need for ***better coverage on language, culture, countries, user groups, and domain***, as suggested by Figure 7 and LLMs' worse performance on stance detection in economic domain and culturally aware language inference. This will help models better generalize across populations and social contexts. We need large and diverse datasets that can reflect the linguistic and cultural diversity of different user groups, which ensures that the models recognize and respond to social cue variations specific to different communities. Consider standard languages or mainstream user groups, versus low-resource languages and dialects, and vulnerable populations such as older adults or people with cognitive impairments.

5.2 Recommendations for Data Structure

There is a need for ***higher interactivity in both data and evaluation***. The field has an abundance of static resources and fewer interactive datasets. Our results show that dialogue context improves performance, and others have also argued that without interactivity, language models may be unable to fully develop key pillars of social intelligence, including Theory of Mind Sap et al. (2022); Bender and Koller (2020). In this interactive settings, there will be an opportunity for a reduced focus on performance metrics like accuracy, and an increased focus on explainability, with socially intelligent AI systems that understand and can explain the factors that underlie their behaviors.

There can be significant shifts in people's values, beliefs and perspectives over time with societal changes like social movements, generational shifts and globalization. As a result, what social intelligence entails is constantly evolving. Thus, ***data should undergo dynamic evolution*** to accurately capture social intelligence over different timeframes. Researchers can also consider a dynamic and flexible framework to allow future customization and extension (Zhou et al., 2023a).

Additionally, humans communicate using various modalities beyond language, such as gestures and facial expressions. Future datasets are encouraged to ***incorporate multiple modalities*** to help AI systems develop a more accurate and well-rounded understanding of social contexts and social cues, leading to increased social intelligence.

5.3 Recommendations for Data Collection

Historically, social AI datasets have drawn heavily on randomly sourced crowdworkers who annotate datasets that have been scraped from social media or other online sources. There are at least three reasons why this paradigm will need to be replaced.

The first concern is the issue of representation. A random sample of crowdworkers may not contain a fair representation of diverse viewpoints from a wide variety of sociodemographic backgrounds. Similar biases appear in randomly sampled social media data. Representation should extend beyond nationality to include diverse local regions, vulnerable populations, and people of different ages and genders. Both annotation and evaluation criteria should be designed in a way that ***accounts for sociolinguistic variation*** and ***considers diverging perspectives***. Relatedly, crowdworkers may not be equipped to consistently identify subtle social cues. For this reason, we support ***increasingly interdisciplinary, expert annotation efforts*** in which domain experts such as linguists, psychologists, anthropologists, and sociologists work to annotate high-quality social AI data resources.

Second, by passively observing decontextualized data, annotators may be unable to fully understand the social context behind any observed behaviors. There may be frequent misalignment between the behaviors expressed in random internet data and the lived experiences of the annotators. This motivates a more ***active paradigm of dataset construction*** in which annotators *participate* in the social interaction, and are thus *de facto* experts on its situational context, any operational norms and cultural expectations that govern their behavior, as well as their own cognitive factors like personally motivating beliefs, intents, and emotions. As an added benefit, such active construction will naturally produce data with a high degree of interactivity (see §5.2).

Third, we encourage the field to closely consider how to effectively leverage LLMs to create ***human-in-the-loop collaborative datasets***, which applies both to the active generation of data previously mentioned and co-annotation of other social constructs (Li et al., 2023). Note that this differs from using LLMs to simulate synthetic social interaction data. In fact, we argue that it is still unclear whether simulation can produce high quality data with practical validity, since recent studies have shown caricatures and stereotypes in LLM-based simulations (Durmus et al., 2023; Cheng et al., 2023).

Last but not least, we call for the development of annotation tools to facilitate the collection, visualization and annotation of different constructs in social intelligence, to allow for easy plug-in to existing crowdsourcing platforms and to support reproducible data collection.

5.4 Recommendations for Data Ethics

Social AI datasets must be designed with ethical considerations, such as fairness, transparency, and privacy, to avoid perpetuating stereotypes or biases, and to respect user privacy. We envision that social AI dataset construction takes a community-centric approach where domain users co-design the tasks and data collection efforts with researchers (i.e., tasks of the community, by the community, and for the community), in addition to interdisciplinary collaboration among research fields. This process will also benefit from protocols, compliance guides, culture- or country specific data use agreements to address any legal and ethical issues for creating and maintaining social AI datasets.

6 Conclusions

We introduce *Social AI Data infrastructure* with a theoretically grounded taxonomy and a data library of 480 NLP datasets, which facilitates standardization of the social intelligence concept in AI systems and organization of previous NLP datasets. We also conduct comprehensive analysis on the data library and evaluate LLMs' performance, offering insights on the current data landscape and future dataset development to advance social intelligence in NLP systems. It enables curation of high-quality datasets and holistic development of social intelligence in the NLP field.

Limitations

Although we try to be comprehensive, the datasets in our data library are not exhaustive as it is practically impossible to capture all datasets on socially intelligence. Moreover, we only crawled datasets on ACL Anthology, which is not representative of the whole academic space. As such, our analysis is more on relative comparison in the NLP domain rather than interpretation of the absolute figure. We encourage future work to further extend and contribute to our initial data library. Moreover, since LLMs have been trained on a large number of data, there may be data leakage issue where LLMs have seen some datasets in our experiment, making the performance reported higher than their actual capability. On top of that, since our work focuses more on obtaining insights about future dataset design (data aspect) instead of testing LLMs' social capability comprehensively (model aspect), we only select one simple and one nuanced dataset for each category for comparison purposes. Future work could leverage upon our infrastructure to design a comprehensive evaluation set for social intelligence to get insights on how models perform along each dimension of social intelligence.

Ethical Statement

This study has been approved by the Institutional Review Board (IRB) at the researchers' institution, and we obtained participant consent with a standard institutional consent form. One ethical concern is that models will become more capable of undesirable outcomes like persuasive misinformation or psychological manipulation as they become more socially intelligent. There may also be concerns that skilled anthropomorphic models will come to replace humans. These can not only lead to loss of trust in users (Mori et al., 2012) but also harm users' well-being (Salles et al., 2020). Our work proposes a standard concept and analysis the landscape and these risks are beyond the scope, but we acknowledge their presence and encourage future social AI data and systems to have clearer guidelines on the capabilities and limitations of AI systems to prevent deceptive and manipulative behaviours when advancing social intelligence.

Acknowledgements

We are thankful to the members from SALT Lab for their helpful feedback. Minzhi Li is supported by the A*STAR Computing and Information Science (ACIS) Scholarship. Caleb Ziems is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655. This work was partially sponsored by NSF IIS-2247357 and IIS-2308994.

References

Khalid Alnajjar and Mika Hämäläinen. 2021. ! qu\'e maravilla! multimodal sarcasm detection in spanish: a dataset and a baseline. *arXiv preprint arXiv:2105.05542*.

Plamen P Angelov, Eduardo A Soares, Richard Jiang, Nicholas I Arnold, and Peter M Atkinson. 2021. Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 11(5):e1424.

Ian Apperly. 2010. *Mindreaders: the cognitive basis of" theory of mind"*. Psychology Press.

Mohamed Jehad Baeth. 2019. *Provenance use in social media software to develop methodologies for detection of information pollution*. Ph.D. thesis.

Albert Bandura. 2009. Social cognitive theory of mass communication. In *Media effects*, pages 110–140. Routledge.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. Mindcraft: Theory of mind modeling for situated dialogue in collaborative tasks. *arXiv preprint arXiv:2109.06275*.

Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. 2020. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.

Michael L Barnes and Robert J Sternberg. 1989. Social intelligence and decoding of nonverbal cues. *Intelligence*, 13(3):263–287.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Nicolas Bertagnolli. 2020. Counsel chat: Bootstrapping high-quality therapy data.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In *Proceedings of the second workshop on computational modeling of people's opinions, personality, and emotions in social media*, pages 36–41.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020a. A sentiment analysis dataset for code-mixed malayalam-english. *arXiv preprint arXiv:2006.00210*.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed tamil-english text. *arXiv preprint arXiv:2006.00206*.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. Casino: A corpus of campsite negotiation dialogues for automatic negotiation systems. *arXiv preprint arXiv:2103.15721*.

Kushal Chawla, Weiyan Shi, Jingwen Zhang, Gale Lucas, Zhou Yu, and Jonathan Gratch. 2023. Social influence dialogue systems: A survey of datasets and models for social influence tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 750–766.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do llms understand social knowledge? evaluating the sociability of large language models with socket benchmark.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Marco Cristani, Giulia Paggetti, Alessandro Vinciarelli, Loris Bazzani, Gloria Menegaz, and Vittorio Murino. 2011. Towards computational proxemics: Inferring social relations from interpersonal distances. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 290–297. IEEE.

Alan Cruse. 2004. Meaning in language: An introduction to semantics and pragmatics.

Marco Del Tredici, Diego Marcheggiani, Sabine Schulte im Walde, and Raquel Fernández. 2019. You shall know a user by the company it keeps: Dynamic representations for social media users in nlp. *arXiv preprint arXiv:1909.00412*.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Daantje Derks, Arjan ER Bos, and Jasper Von Grumbkow. 2007. Emoticons and social interaction on the internet: the importance of social context. *Computers in human behavior*, 23(1):842–849.

JA DeVito. 2016. The interpersonal communication book, global edition.

Esin Durmus, Karina Nyugen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Mica R Endsley. 1990. *Situation awareness in dynamic human decision making: Theory and measurement*. Ph.D. thesis, University of Southern California Los Angeles, CA.

Thomas Erickson. 2009. 'social'systems: designing digital systems that support social intelligence. *Ai & Society*, 23(2):147–166.

Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. 2022. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2):144–160.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. *arXiv preprint arXiv:2011.00620*.

Martin E Ford and Marie S Tisak. 1983. A further search for social intelligence. *Journal of Educational Psychology*, 75(2):196.

Simona Frenda, Alessandro Pedrani, Valerio Basile, Soda Marem Lo, Alessandra Teresa Cignarella, Raffaella Panizzon, Cristina Sánchez-Marco, Bianca Scarlini, Viviana Patti, Cristina Bosco, et al. 2023. Epic: Multi-perspective annotation of a corpus of irony. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13844–13857.

Omar Juárez Gambino, Hiram Calvo, and Consuelo-Varinia García-Mendoza. 2018. Distribution of emotional reactions to news articles in twitter. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Deepanway Ghosal, Siqi Shen, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2022. Cicero: A dataset for contextualized commonsense inference in dialogues. *arXiv preprint arXiv:2203.13926*.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599.

Edward Twitchell Hall and Mildred Reed Hall. 1987. Hidden differences: Doing business with the japanese. *(No Title)*.

William Held, Camille Harris, Michael Best, and Diyi Yang. 2023. A material lens on coloniality in nlp. *arXiv preprint arXiv:2311.08391*.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 588–602.

David M Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, pages 169–182. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Thelma Hunt. 1928. The measurement of social intelligence. *Journal of Applied Psychology*, 12(3):317.

Anthony Hunter, Lisa Chalaguine, Tomasz Czernuszenko, Emmanuel Hadoux, and Sylwia Polberg. 2019. Towards computational persuasion via natural language argumentation dialogues. In *KI 2019: Advances in Artificial Intelligence: 42nd German Conference on AI, Kassel, Germany, September 23–26, 2019, Proceedings 42*, pages 18–33. Springer.

Zunaira Jamil. 2017. *Monitoring tweets for depression to detect at-risk users*. Ph.D. thesis, Université d'Ottawa/University of Ottawa.

Pratik Jayarao and Aman Srivastava. 2018. Intent detection for code-mix utterances in task oriented dialogue systems. In *2018 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, pages 583–587. IEEE.

Yuru Jiang, Yang Xu, Yuhang Zhan, Weikai He, Yilin Wang, Zixuan Xi, Meiyun Wang, Xinyu Li, Yu Li, and Yanchao Yu. 2022. The crecil corpus: a new dataset for extraction of relations between characters in chinese multi-party dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2337–2344.

Zhijing Jin, Sydney Levine, Fernando Gonzalez Adauto, Ojasv Kamal, Maarten Sap, Mrinmaya Sachan, Rada Mihalcea, Josh Tenenbaum, and Bernhard Schölkopf. 2022. When to make exceptions: Exploring language models as accounts of human moral judgment. *Advances in neural information processing systems*, 35:28458–28473.

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *12th International Conference on Language Resources and Evaluation, LREC 2020*, pages 1008–1018. European Language Resources Association (ELRA).

John F Kihlstrom and Nancy Cantor. 2000. Social intelligence.

Bennett Kleinberg, Isabelle Van Der Vegt, and Maximilian Mozes. 2020. Measuring emotions in the covid-19 real world worry dataset. *arXiv preprint arXiv:2004.04225*.

Corinne Kosmitzki and Oliver P John. 1993. The implicit use of explicit conceptions of social intelligence. *Personality and individual differences*, 15(1):11–23.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in Artificial Intelligence*, 5:62.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5872–5877.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022. Personachatgen: Generating personalized dialogues using gpt-3. In *Proceedings of the 1st Workshop on Customized Chat Grounding Persona and Knowledge*, pages 29–48.

Minzhi Li, Taiwei Shi, Caleb Ziems, Min-Yen Kan, Nancy F Chen, Zhengyuan Liu, and Diyi Yang. 2023. Coannotating: Uncertainty-guided work allocation between human and large language models for data annotation. *arXiv preprint arXiv:2310.15638*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

Q Vera Liao and S Shyam Sundar. 2022. Designing for responsible trust in ai systems: A communication perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1257–1268.

Filip Lievens and David Chan. 2017. Practical intelligence, emotional intelligence, and social intelligence. *Handbook of employee selection*, pages 342–364.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

Eliot Maës, Thierry Legou, Leonor Becerra, and Philippe Blache. 2023. Studying common ground instantiation using audio, video and brain behaviours: the brainkt corpus. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 691–702.

Herbert A Marlowe. 1986. Social intelligence: Evidence for multidimensionality and construct independence. *Journal of educational psychology*, 78(1):52.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Masahiro Mori, Karl F MacDorman, and Norri Kageki. 2012. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100.

Fred August Moss and Thelma Hunt. 1927. Are you socially intelligent? *Scientific American*, 137(2):108–110.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. *arXiv preprint arXiv:2011.01589*.

Silviu Oprea and Walid Magdy. 2019. isarcasm: A dataset of intended sarcasm. *arXiv preprint arXiv:1911.03123*.

Baolin Peng, Chunyuan Li, Zhu Zhang, Chenguang Zhu, Jinchao Li, and Jianfeng Gao. 2020. Raddle: An evaluation benchmark and analysis platform for robust task-oriented dialog systems. *arXiv preprint arXiv:2012.14666*.

Xianglan Peng. 2021. Research on emotion recognition based on deep learning for mental health. *Informatica*, 45(1).

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Bahar Radfar, Karthik Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 959–963.

Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2018. Semi-supervised user geolocation via graph convolutional networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2009–2019, Melbourne, Australia. Association for Computational Linguistics.

S Ramaneswaran, Sanchit Vijay, and Kathiravan Srinivasan. 2022. Tamilatis: dataset for task-oriented dialog in tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 25–32.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Shaurya Rohatgi, Yanxia Qin, Benjamin Aw, Niranjana Unnithan, and Min-Yen Kan. 2023. The ACL OCL corpus: Advancing open science in computational linguistics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10348–10361, Singapore. Association for Computational Linguistics.

Jonathan P Roiser and Barbara J Sahakian. 2013. Hot and cold cognition in depression. *CNS spectrums*, 18(3):139–149.

Tessa Rusch, Saurabh Steixner-Kumar, Prashant Doshi, Michael Spezio, and Jan Gläscher. 2020. Theory of mind and decision science: Towards a typology of tasks and computational models. *Neuropsychologia*, 146:107488.

Arleen Salles, Kathinka Evers, and Michele Farisco. 2020. Anthropomorphism in ai. *AJOB neuroscience*, 11(2):88–95.

Wesley Santos, Amanda Funabashi, and Ivandré Paraboni. 2020. Searching brazilian twitter for signs of mental health issues. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6111–6117.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019a. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *arXiv preprint arXiv:2210.13312*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019b. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Poorvi Shetty. 2023. Poorvi@ dravidianlangtech: Sentiment analysis on code-mixed tulu and tamil corpus. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 124–132.

David Silvera, Monica Martinussen, and Tove I Dahl. 2001. The tromsø social intelligence scale, a self-report measure of social intelligence. *Scandinavian journal of psychology*, 42(4):313–319.

Sonali Singh and Navita Srivastava. 2023. Emotion recognition for mental health prediction using ai techniques: An overview. *International Journal of Advanced Research in Computer Science*, 14(3).

Parinaz Sobhani, Diana Inkpen, and Xiaodan Zhu. 2017. A dataset for multi-target stance detection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 551–557.

Michael M Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of depression-related posts in reddit social media forum. *Ieee Access*, 7:44883–44893.

Chenhao Tan and Lillian Lee. 2014. A corpus of sentence-level revisions in academic writing: A step towards understanding statement strength in communication. *arXiv preprint arXiv:1405.1439*.

Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. 2023. Depac: a corpus for depression and anxiety detection from speech. *arXiv preprint arXiv:2306.12443*.

Edward L Thorndike. 1921. Intelligence and its measurement: A symposium–i. *Journal of Educational psychology*, 12(3):124.

Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2021. Pride: Predicting relationships in conversations. In *The Conference on Empirical Methods in Natural Language Processing*, pages 4636–4650. ACL.

John C Turner. 1991. *Social influence.* Thomson Brooks/Cole Publishing Co.

Juan Vásquez, Scott Andersen, Gemma Bel-Enguix, Helena Gómez-Adorno, and Sergio-Luis Ojeda-Trueba. 2023. Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 202–214.

Philip E Vernon. 1933. Some characteristics of the good judge of personality. *The Journal of Social Psychology*, 4(1):42–57.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Daniela Wawra. 2013. Social intelligence: The key to intercultural communication. In *Intercultural Negotiations*, pages 29–42. Routledge.

Eugene A Weinstein. 1969. The development of interpersonal competence. *Handbook of socialization theory and research*, pages 753–775.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

Yirui Wu, Hao Li, Lilai Zhang, Chen Dong, Qian Huang, and Shaohua Wan. 2023. Joint intent detection model for task-oriented human-computer dialogue system using asynchronous training. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(5):1–17.

Yuwei Wu, Xuezhe Ma, and Diyi Yang. 2021. Personalized response generation via generative split memory network. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1956–1970.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018a. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018b. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Zheng Zhang, Ryuichi Takanobu, Qi Zhu, MinLie Huang, and XiaoYan Zhu. 2020. Recent advances and challenges in task-oriented dialog systems. *Science China Technological Sciences*, 63(10):2011–2027.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023a. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023b. Cobra frames: Contextual reasoning about effects and harms of offensive statements. *arXiv preprint arXiv:2306.01985*.

Qi Zhu, Kaili Huang, Zheng Zhang, Xiaoyan Zhu, and Minlie Huang. 2020. Crosswoz: A large-scale chinese cross-domain task-oriented dialogue dataset. *Transactions of the Association for Computational Linguistics*, 8:281–295.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023a. NormBank: A knowledge bank of situational social norms. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7756–7776, Toronto, Canada. Association for Computational Linguistics.

Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023b. Normbank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008*.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2023c. Can large language models transform computational social science? *arXiv preprint arXiv:2305.03514*.

A Other Analysis

Perspective-Taking Perspective-taking is the act of considering an alternative point of view for the same situation, which is one aspect of social intelligence (Kosmitzki and John, 1993). Some work starts to pay more attention to two different perspectives which are ***intended and perceived point of views***. For example, Oprea and Magdy (2019) points out the difference between intended and perceived sarcasm from the perspectives of the author and audience, which is often overlooked in previous work. Thus, they asked for self-reported annotations to capture the intended sarcasm. The same also holds for other factors like emotion, which can be intended emotion by the author (Kleinberg et al., 2020) or aroused emotion among the audience (Gambino et al., 2018).

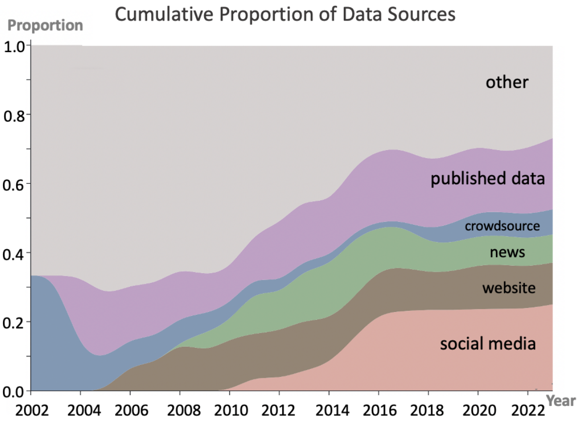

Figure 6: Percentage for major sources (social media, website, news, crowdsourcing, existing datasets) and other sources (e.g. book, speech etc.) over time.

Distribution of Data Sources Most datasets in our data library use ***social media*** as sources of data (see Figure 6). The prevalence of data collection from social media has experienced ***a significant surge from 2010***. This might be due to an increase in the use of Twitter data (Baeth, 2019). In the meantime, relative proportions of traditional media like news and websites has experienced a decrease since then.

The second popular data source is ***previously built data resources***. New datasets leverage and extend upon previous ones in cases like translating to low-resource language (Ramaneswaran et al., 2022), introducing new evaluation criteria (Peng et al., 2020) and adding new layers of annotation (Tigunova et al., 2021).

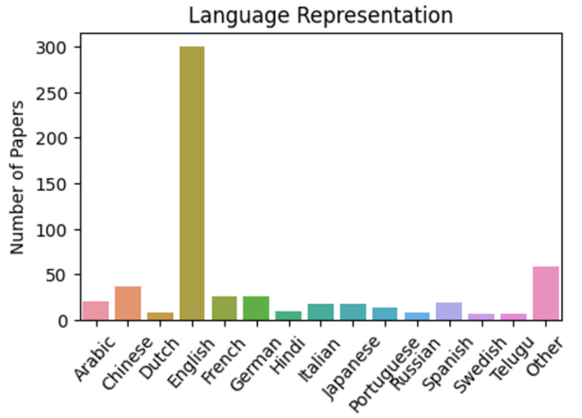

Figure 7: Distribution of datasets in different languages.

Language and Culture Representation Datasets surveyed in our data library covered up to 49 different types of languages. Figure 7 shows that ***majority of study (62.5%) uses English data*** to explore social intelligence and the number of such work is much higher than those in other languages. Moreover, there are more recent research efforts on code-mixing datasets about social intelligence, suggesting an increased representation of multilingual community (Chakravarthi et al., 2020a,b; Shetty, 2023).

Additionally, most datasets in the data library (97.9%) has unspecified cultural representation. However, the same sentence could have different meanings under different cultural contexts as social interpretations and social interactions vary from culture to culture. Therefore, there is a strong need for more future datasets with generations and annotations from different cultural backgrounds.

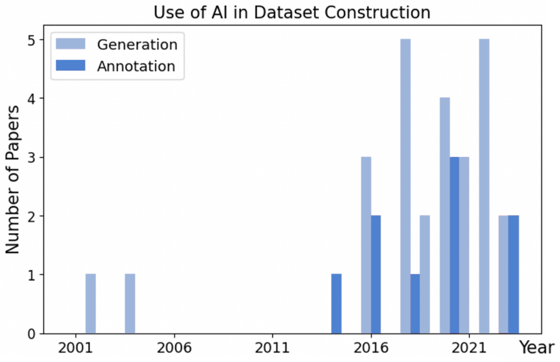

Figure 8: Number of papers that used AI for dataset generation or annotation.

Overall, most of datasets contain textual content that is purely human-generated (95.0%) and manually labeled (98.1%). From Figure 8, we can see there is an ***increasing trend in adoption of AI*** for generating and annotating datasets related to social intelligence. We can also see that number of work ***using AI for generation is more than those for annotation***.

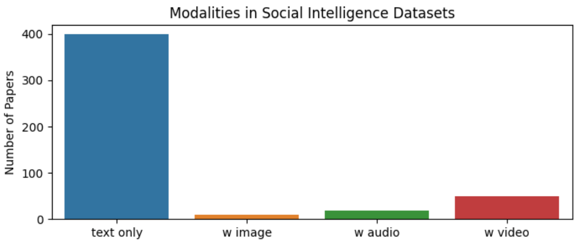

Figure 9: Distribution of modalities in datasets on social intelligence being surveyed.

Incorporation of Different Modalities Because we only use crawled data from ACL Anthology, the majority datasets on social intelligence we surveyed are only in textual format. However, different modalities like image, audio and video can enhance learning of social intelligence with enriched social information embedded in other modalities.

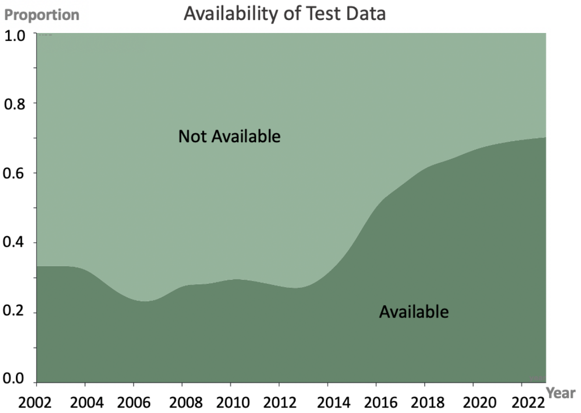

Figure 10: Proportion of available and unavailable social intelligence dataset (test set) over the years.

More Open-sourced Community Figure 10 shows a promising trend where proportion of social intelligence datasets that are made available increases over the years. From 2014 onwards, the proportion of available has experienced a surge and there are ***more publicly available social intelligence datasets*** than unavailable ones from 2016 onwards. Unavailable social intelligence datasets usually contain sensitive information such as mental illness (Santos et al., 2020; Tasnim et al., 2023) and sexual orientation (Vásquez et al., 2023). Additional measures like anonymizing, dara encryption and access control should be in place to protect data confidentiality while ensuring future work could have secure ways to use these data to advance research in sensitive domains.

B LLM Inference

Model We run inferences using Claude-v2, GPT-4-1106 and Llama2-13b models.

Prompt We follow the recommended practice of prompting as described by Ziems et al. (2023c). We design prompts in MCQ format, give instructions after the context and clarify specific social concepts when necessary.

Temperature Setting We set temperature to 0 for classification tasks to ensure consistency and 0.7 for generation tasks to allow some diversity.

C Details of Human Evaluation

Dataset	Criteria
DailyDialogue	information content, appropriateness, engagement, naturalness, human-likeness
PersonaChat	information content, appropriateness, engagement, naturalness, human-likeness, persona consistency
Positive Reframing	meaning preservation, degree of positivity, naturalness, human-likeness
Counsel Chat	information content, appropriateness, engagement, naturalness, human-likeness, empathy
Convincing Arguments	persuasiveness, information content naturalness, human-likeness
PersuasionforGood	information content, appropriateness, engagement, naturalness, human-likeness, persuasiveness

Table 2: Criteria for different datasets used in human evaluation. They include general criteria like appropriateness and information content (Howcroft et al., 2020) and task-specific criteria such as meaning preservation in positive reframing and persona consistency in persona controlled dialogue.

For interpretability, we also provide different criteria (see Table 2) for each task and collect human's free-text explanation on top of an overall judgment. Below is the qualitative analysis for their free-text inputs for each dataset:

Daily Dialogue A better responses is more specific and thoughtful and goes beyond the straightforward question a bit to make a personal anecdote. Some machine generated responses are too artificially friendly and sympathetic.

PersonaChat A response is better if it expresses more information (information content), responds directly to the question (appropriateness, naturalness) and expands on simple facts with personal details (naturalness, engagement).

Positive Reframing A better response is more *multifaceted* and *nuanced* by acknowledging all aspects of the input text. It better acknowledges and preserves the original meaning. Some machine generated responses feel forced and rated as worse.

Counsel Chat Better responses are more detailed, address the complexity of the situation and provide more specific and actionable advice.

ArgumentsPairs Better arguments assert claims with evidence, provides specificity and contextualization of evidence and offer multiple complementary ideas that address different sub-concerns.

PersuasionforGood Better responses are more specific and have a clear call to action. They better express emotional (pathos) and logical (logos) appeals. They take a *genuine interest* in the listener.

In addition, we also collect crowdworkers' perceptions and comments on whether they think the text is generated by machine or human.

We find that majority of people cannot correctly identify which response is generated by machine and a significant proportion choose the option *'both generations are produced by human'*. Thus, more regulation and transparency on who generates the texts are needed since it is hard for people to distinguish human and machine generated texts.

In general, they perceive more specific and thoughtful answers to be human generated and those brief and incomplete answers to be machine generated. Occasionally, some people hold an opposite view that more detailed and structured responses are from machine. This suggests for generation tasks, LLM has exceeded average human performance (see Table 1) and it has also been acknowledged in *human's perception* for a certain proportion of population.

D Keywords

E Dataset Selection Criteria

We provide justifications for the choice of one simpler and one challenging dataset used for LLM probing for each subcategory.

	Title	Abstract
Dataset Keywords	’benchmark’, ’corpus’, ’dataset’ , ’annotat’	’introduce/build a dataset/benchmark’, ’introduce/build a largescale dataset/benchmark’, ’introduce/build a large-scale dataset/benchmark’, ’the first benchmark’, ’the first largescale benchmark’, ’the first large-scale benchmark’
Social Intelligence Keywords	‘cognitive’,‘sentiment’,‘agreement’,‘debate’,‘bargain’, ‘commonsense’,‘emotion’, ‘polar’, ‘expressive’, ‘argument’, ‘autistic’, ‘information-providing’, ‘interaction’, ‘opinion’,‘negotiation’, ‘dialog’, ‘affect’, ‘public speaking’, ‘semeval’,‘stereotype’, ‘hate speech’, ‘stance’, ‘persona’, ‘conversation’,‘communicative’,‘communicate’, ‘recommendation’, ‘intent’, ‘communication’, ‘gender’, ‘age’, ‘emotional’, ‘sympathy’,‘empathy’, ‘mental’, ‘norm’, ‘culture/cultural’, ‘social relation’, ‘speaker’, ‘author profiling’, ‘moral’, ‘ethic’, ‘privacy/secret’, ‘socially aware’, ‘prosocial’, ‘moral’, ‘social situation’, ‘social context’, ‘social commonsense’, ‘style’, ‘depression’, ‘anxiety’, ‘persuasion’, ‘recommendation’, ‘theory of mind’, ‘audience’, ‘chat’	

Table 3: Keywords to filter for papers on dataset collection and social intelligence.

Intent SNIPs dataset is about query intent classification while the iSarasm dataset involves identification of sarcastic intents. The latter is more challenging as it demands understanding of contextual complexity and language ambiguity which makes the task more nuanced.

Belief SemEval-Task6 is about stance detection for texts in the political domain, which is the dominant domain in stance detection. WTWT is a more challenging dataset as it consists of data from the economic domain, with limited previous data efforts in the area of stance detection.

Emotion GoEmotions dataset is more challenging as it has more fine-grained emotion classes (28 classes) than SemEval Task 1 (11 classes).

Social Situation CICERO dataset is more challenging as the social situation is implicitly reflected in a dialogue compared to the SocialIQa dataset where the social situation is explicitly described.

Social Norm NormBank contains mostly common scenarios in daily life and MoralExceptQA is more challenging as the situations include cases with exceptions for moral judgment (e.g. permissibility of cutting queue which is usually unacceptable in a really urgent context).

Chitchat PersonaChat is more challenging than DailyDialogue as it has additional persona constraints for dialogue response generation.

Persuasion and Therapy The less challenging datasets (ConvincingArguments and Positive Reframing datasets) contain static settings with single-turn generation whereas the more challenging ones (PersuasionforGood and CounselingChat) are for interactive multi-turn settings where more contextual information needs to be taken into account during response generation.

F Annotator Qualification

For selection criteria of human annotators, we require an accepted number of tasks higher than 500 and an approval rate higher than 98%. Crowdworkers who met this criteria will take a qualification round which contains exactly the same questions in the actual round. Their submissions for the qualification round were manually verified by the authors and qualified workers were given access to the actual round.