Situational Evaluation for Social Intelligence of Large Language Models

Anonymous ACL submission

Abstract

The academic intelligence of large language models (LLMs) has made remarkable progress in recent times, but their social intelligence performance remains unclear. Inspired by established human social intelligence frameworks, particularly Daniel Goleman's social intelligence theory, we have developed a standardized social intelligence test based on real-world social scenarios to comprehensively assess the social intelligence of LLMs, termed as the Situational Evaluation for Social Intelligence (SESI). We conducted an extensive evaluation with 13 popular and state-of-art LLMs on SESI. The results indicate the social intelligence of LLMs still has significant room for improvement, with superficially friendliness as a primary reason for errors. Moreover, there exists a relatively low correlation between the social intelligence and academic intelligence exhibited by LLMs, suggesting that social intelligence is distinct from academic intelligence for LLMs. Additionally, while it is observed that LLMs can't "understand" what social intelligence is, their social intelligence, similar to that of humans, is influenced by social factors.

1 Introduction

The ability to understand and manage social relationships is one fundamental dimension of human intelligence, commonly denoted as social intelligence (Thorndike, 1920). Social intelligence enables humans to reduce conflicts and foster cooperation, thus navigating the social world. It not only correlates closely with individual success and life satisfaction (Joseph and Lakshmi, 2010; Zakirova and Frolova, 2014), but also is one of the most important ingredients in humans' survival as a species in the long run (Albrecht, 2006).

As a core component of human intelligence, social intelligence stands as an indispensable milestone on the path to achieving artificial general intelligence (AGI) (Sterelny, 2007). On one hand, social intelligence is necessary for facilitating effective communication and collaboration both among artifacts and between artifacts and humans (Dautenhahn, 1995). On the other hand, social intelligence provides the foundation to deeply learn for AI systems, particularly large language models (LLMs), as language is inherently social, and meaning is constructed through social interactions (Wittgenstein, 2019). Moreover, social intelligence is closely associated with crucial issues of AI alignment and governance. Individuals with high social intelligence can effectively manage conflicts between individual and group objectives (Korinek and Balwit, 2022) and avoid toxic behaviors by equipping awareness of the impact on others (Albrecht, 2006).

While the importance of social intelligence is widely acknowledged (Hovy and Yang, 2021), evaluating it within recently developed advanced AI systems, particularly LLMs such as ChatGPT (OpenAI, 2021, 2023), Claude (Anthropic, 2023), and LLaMA (Touvron et al., 2023a,b), remains limited. Current research primarily examines the academic intelligence of LLMs, highlighting their proficiency in social isolated tasks like tool use, automated theorem proving and so on (Chang et al., 2023; Sarkisyan et al., 2023), while the social intelligence of LLMs, crucial for real-world applications, is often perceived as a "side effect" and has not been comprehensively established in a robust manner. Some researchers assess the social intelligence of LLMs based on classic tests of human social intelligence, such as ToMi (Le et al., 2019) and Faux-Pas (Shapira et al., 2023b). These well-established tests have a long history, making it likely that LLMs have been exposed to and trained on them (Shapira et al., 2023a). Some other researchers assesses social intelligence of LLMs in the context of social factor understanding, exemplified by datasets such as SocialIQA (Sap et al., 2019), SocKET (Choi et al., 2023) and SECEU (Wang et al., 2023). These

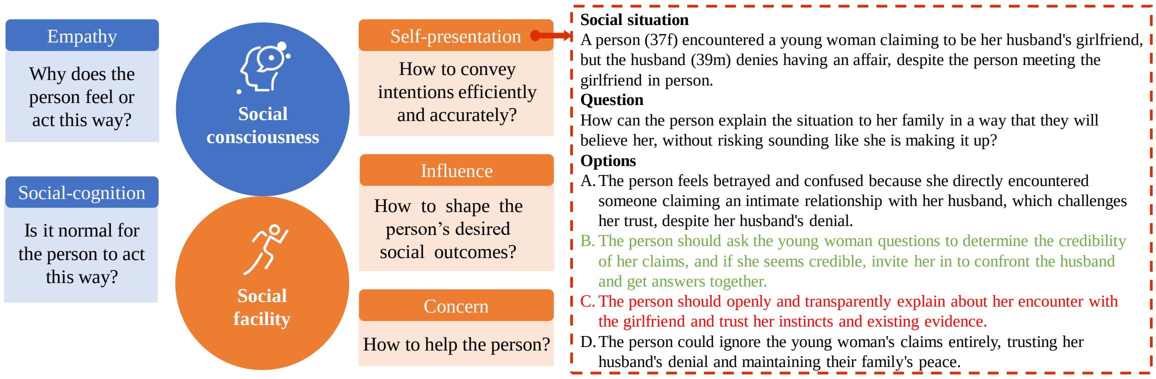

Figure 1: Overview of Situational Evaluation for Social Intelligence (SESI). SESI assesses social intelligence of LLMs from two directions: social awareness and social facility, including five specific social abilities. In the given example, the correct answer and incorrect choice by gpt-3.5-turbo are highlighted.

datasets focus on assessment of social awareness, the ability to comprehend and track agents' inner states, such as emotions, beliefs, motivations and so on, while ignoring social facility, the ability to act smoothly and efficiently in relationships, which is necessary to guarantee fruitful interactions. There are also two innovative benchmarks, SOTOPIA (Zhou et al., 2023) and EmoBench (Sabour et al., 2024). However, they either employ manually crafted social contexts and goals, introducing subtle differences from real-world interactive scenarios, or solely focus on a single social factor, thereby limiting the ability to comprehensively assess social intelligence. Therefore, there is a need for a dynamic and comprehensive benchmark to go beyond existing benchmarks, in order to fully assess the social intelligence of LLMs.

To fill the gap, we develop the Situational Evaluation for Social Intelligence (SESI), a comprehensive and challenging benchmark for assessing LLMs' social intelligence in real and complex social situations, as shown in Figure 1. SESI contains 500 test items, each of which consist of a social situtaion-question pair and four comments that seem to offer alternative explanations. Specifically, the social situations and questions are derived from authentic requests for assistance posted by users on Reddit Relationships community[1], and the correct answers are determined based on the most endorsed responses, which reflect group consensus (Petrides, 2011; Weis, 2008). Compared to the previously mentioned benchmarks, SESI possesses two distinctive advantages: 1) comprehensive. SESI is grounded in established human social intelligence frameworks, including Daniel Goleman's social intelligence theory (Daniel, 2006) and S.P.A.C.E theory (Albrecht, 2006), thus comprehensively assessing all social skills. 2) Dynamic. Test items in SESI can be automatically generated based on Reddit Q&A posts. This allows for automatic updates over time, representing a core distinction from previous evaluations conducted on static datasets.

We then conducted an evaluation of a spectrum of mainstream and widely-adopted LLMs on SESI, and obtained the following findings: 1) The social intelligence of LLMs still has significant room for improvement, as evidenced by the best-performing model, gpt-3.5-turbo-0613, which achieves only 55.2% performance. 2) The social intelligence of LLMs is distinct from academic intelligence, warranting investigation as a separate form of intelligence. 3) LLMs are superficially friendly, following fixed friendly patterns without grounding them in real social situations, which is the main reason for the errors made by LLMs in social judgments. 4) Social intelligence of LLMs, similar to that of human beings, is influenced by social factors, including personality, gender, social role and person.

2 SESI: The Situational Evaluation for Social Intelligence

2.1 Introduction to SESI

Aligned with established human social intelligence frameworks (Daniel, 2006; Albrecht, 2006), we have developed a standardized test for assessing social intelligence in LLM agents, termed as the Situational Evaluation for Social Intelligence (SESI). SESI is designed to evaluate two core components of social intelligence: social consciousness, which deals with feelings towards others, and social facility, which is the behavioral manifestations in possession of consciousness (Details in Section 2.2). SESI draws inspiration from real-life social scenar-

[1] https://www.reddit.com/r/relationships/

ios, with each test item comprising a social situation, a contextual question and four options that seem to offer alternative explanations. To elaborate, the social situations depict interpersonal relationships and entanglements in social events involving a central figure, "the person." The questions inquire about potential resolutions to the challenges faced by "the person" within the given social context. The four response options offer varied inferences related to the scenario. LLM agents are required to comprehend the social context and make inferences to select the most appropriate, intelligent, or logically sound comment from the provided options.

2.2 Social intelligence components in SESI

The SESI assesses LLMs' proficiency in five social abilities, defined below.

- Social Consciousness: This pertains to the ability to comprehend others and social situations. It includes the following aspects:
  - Empathy: The ability to comprehend and infer the thoughts, feelings, and intentions of others within a given context.
  - Social Cognition: The ability to understand complex social situations, such as why a particular situation is awkward.
- Social Facility: This encompasses the ability to act smoothly and efficiently in interpersonal relationships. It includes the following aspects:
  - Self-presentation: The ability to convey intentions efficiently and accurately.
  - Influence: The ability to shape desired social outcomes, typically involving altering others' perspectives.
  - Concern: The ability to identify others' needs and take appropriate actions to address them.

2.3 The development of SESI

2.3.1 Social contexts and questions collection

In order to construct SESI, we gathered social contexts and questions from the Reddit Relationships community[1], a forum where users seek advice based on real-world interpersonal interactions. The community comprises 3.4 million members and is dedicated to assisting individuals by providing a platform for interpersonal relationship advice among Redditors. Posters are required to articulate

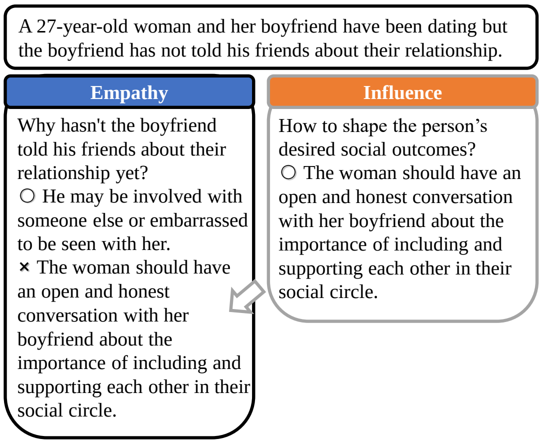

Figure 2: Question-switching answers are collected as the answers to the wrong question that targets a different social ability.

their age, gender, relationship status, context, and pose specific, clearly formulated questions while avoiding biased language.

To implement this data collection process, we utilized PRAW[2] to scrape the 1000 most popular posts in the Reddit Relationships community in 2023. Subsequently, we utilized the GPT-4 model to summarize these posts into social contexts and associated questions and categorize them into five distinct types of social capabilities, in accordance with the social ability definition provided in section 2.2[3]. Throughout this procedure, we excluded posts with multiple updates and external links to maintain data completeness. Additionally, posts that did not pertain to social abilities or that encompassed multiple social abilities were also omitted. We get 547 questions after this step.

2.3.2 Answer collection

Correct answers were generated based on the most widely accepted responses under each post. Based on the widely adopted group consensus scoring principle in social intelligence testing (Petrides, 2011; Weis, 2008), we posit that the top responses beneath each post, endorsed by thousands of individuals, can be considered as optimal answers within the current societal norms. Specifically, we use GPT-4 to filter responses that contain viable suggestions and are the most upvoted, summarizing them into a single sentence to as the correct answer to a question.

[2] https://praw.readthedocs.io/en/stable/

[3] Prompts in the paper can't be provided at this time due to space constraints, but will be released in the future with code.

Wrong answers We collect two groups of wrong answers, including question-switching answers and reversed answers.

Question-Switching Answers were generated by switching the questions asked about the context, as shown in Figure 2. Specifically, we utilized the GPT-4 model to generate answers corresponding to four other social abilities within the same context. Details of the social abilities and corresponding questions can be found in Section 2.2 and Figure 1.

Reversed Answers were answers that diverge from the standpoint of correct answers. Specifically, we utilized the GPT-4 model to generate two reversed answers for each question, with the objective of introducing greater diversity in social comprehension and behavior while ensuring logical coherence.

2.3.3 QA tuple creation

As the final step of the pipeline, data is consolidated into four-way multiple-choice questions. Each test item contains a context-question pair, a correct answer and three incorrect answers. Of these incorrect answers, one is randomly sampled from four available question-switching answers, and two are reversed answers.

Finally, each test item underwent validation by 3 NLP postgraduates. Items that did not align with correct social abilities, lacked correct answers, or had non-unique correct answers were systematically removed. 47 test items are filtered out.

2.4 Dataset Analysis

In this subsection, we present the main statistics of SESI, as illustrated in Figure 3, revealing distinctive features of our benchmark as follows:

- **Comprehensive and balanced assessment of social intelligence abilities.** Illustrated in Figure 3 (d), SESI extends beyond understanding social contexts (empathy, social-cognition) to changing social situations to achieve characters' social goals (self-presentation, influence, concern), which sets SESI apart from conventional common-sense reasoning benchmarks.

- **Long, complex, and diverse social contexts.** Figure 3 (a) shows that the average length of social contexts in the benchmark is 44.2 words, three times that of Social IQA dataset (Sap et al., 2019). Figure 3 (c) indicates that 50% of social situations in SESI involve three or more active characters, signifying their complexity. Moreover, Figure 3 (e) illustrates the diverse array of social relationship types contained within SESI. These distributions of context length, character numbers, and relationship types underscore the challenging nature of the benchmark.

- **Detailed and specific answers.** Figure 3 (b) illustrates that the average answer length in SESI is 25.8 words, notably exceeding prevalent social common-sense reasoning benchmarks, which typically exhibit average answer lengths ranging from 3.6 to 10.5 words (Sap et al., 2019; Zadeh et al., 2019). This highlights the level of detail in answers within SESI. Furthermore, it can be observed that the length distributions of correct and incorrect answers are similar, suggesting that the benchmark prioritizes response substance over length in model assessments.

3 Experimental Setup

3.1 Language Models

We evaluated 13 mainstream and popular LLMs, including OpenAI GPT series[4][5] (GPT-4, GPT-3.5, text-davinci-001, text-davinci-002, text-davinci-003 and DaVinci), Vicuna (Chiang et al., 2023) (Vicuna-13B, Vicuna-33B), LLaMA 2-Chat (Touvron et al., 2023b) (LLaMA 2-7B-chat, LLaMA 2-13B-chat, LLaMA 2-70B-chat), Mixtral (Jiang et al., 2023) (Mistral 7B, Mixtral 8×7B).

3.2 Baseline Benchmarks

We selected benchmarks that are comprehensive, widely adopted, discriminative, and align well with actual usage experience to assess various capabilities of LLMs as accurately as possible.

- Knowledge, which evaluates LLMs' capability on world knowledge, including Natural Questions[6] (NQ) (Kwiatkowski et al., 2019), and Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2020).

- Reasoning, which measures LLMs' general reasoning capability, including BBH (Suz-

[4]Text-davinci-001/2/3 and DaVinci retired after our experiments.

[5]`https://openai.com/blog/openai-api`

[6]For NQ, we evaluate in the closed-book setting, where only the question is provided, without a context document.

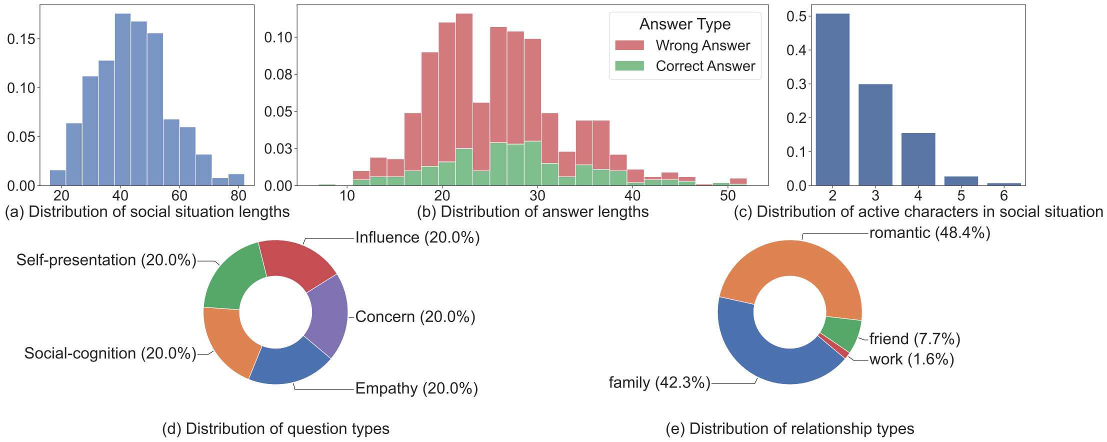

Figure 3: Statistics of SESI benchmark.

gun et al., 2023) and WinoGrande (Sakaguchi et al., 2021).

- Comprehension, which assesses LLMs' capability of reading comprehension, including RACE (Lai et al., 2017) and DROP (Dua et al., 2019).

- Math, which tests mathematical capability, including GSM8K (Cobbe et al., 2021) and MATH (Hendrycks et al., 2021).

- Safety, which scrutinizes LLM's propensity to generate content that is truthful, reliable, non-toxic and non-biased, including TruthfulQA (Lin et al., 2022).

3.3 Evaluation

Prompts. To achieve reliable conclusions, it is crucial to make apples-to-apples LLM comparisons with consistent prompts. For baseline benchmarks, we adopt the identical prompt settings as (Zheng et al., 2023b). For SESI, we refer to the classic Chapin Social Insight Test (Chapin, 1968).

Methods. We adopt a black-box evaluation method. Specifically, when given the test prompt, LLM first generates a free-form response, which is subsequently parsed into the final answer.

Metrics. We default to using the Exact Match (EM) accuracy, except F1 score for DROP dataset.

Hyperparameter. We set temperature to 0.

3.4 Social Factors

A natural question arises: Can the social intelligence of LLMs be controlled, and are the factors shaping human social intelligence transferrable to LLMs? To answer this question, we carefully select five specific social factors for investigation: personality, emotion, gender, social role, and person. These attributes, inspired by prior psychological and sociological research on social intelligence (Goody, 1995; Shafer, 1999; Van der Zee et al., 2002; Spurr and Stopa, 2003; Bilich and Ciarrochi, 2009; Cantor and Kihlstrom, 2013; Dehghanan et al., 2014; Dang, 2014; Mileounis et al., 2015), particularly Daniel's social science theories (Daniel, 2006), can significantly influence the levels of human social intelligence.

Category		Roles
Interpersonal	Family	parent, mother, father, child, son, daughter
	Romatic	partner, husband, wife, girlfriend, boyfriend
	Friend	friend
	Work	coworker, boss, colleague
	School	student, tutor
Occupational	General	saler, teacher, librarian, programmer

Table 1: Roles used in the experiment.

Personality. We choose the widely recognized Big Five personality traits (John et al., 1999) as the fundamental dimensions of personality for our study. Specifically, we incorporated the prompt "You are a/an {personality} individual and score high/low in the trait of {personality} in the Big Five personality traits. This indicates that you are {descriptions}." prior to the basic evaluation prompt. This prompt serves to inform LLM agents of their personality traits.

Emotion. We select three most representative emotions from the classical emotion-performance inverted U-shaped curve (Daniel, 2006), including

Series	Model	Knowledge		Reasoning		Comprehension		Math		Safety	SI
		NQ	MMLU	BBH	WinoGrande	RACE-h	DROP	GSM8K	MATH	TruthfulQA	SESI
GPT	gpt-4-0613	48.6	81.3	84.6	87.1	91.8	87.4	92.1	34.9	79.1	54.4
	gpt-3.5-turbo-0613	38.8	67.4	68.1	55.3	81.2	53.7	76.3	15	61.4	55.2
	text-davinci-003	38.1	63.7	69	70.6	79.5	56.3	59.4	15.6	52.2	38
	text-davinci-002	28.2	62.1	66	65.5	80.5	47.5	47.3	8.5	47.8	42.8
	text-davinci-001	23.5	46.7	38.6	54.6	44.3	33.1	15.6	0	54.2	36.9
	davinci	17.8	34.3	39.1	48	35	16.5	12.1	0	21.4	0.4
LLaMA2	llama-2-70b-chat	40.5	42.5	55.1	58.5	77	58.7	56.9	6	38.3	49.4
	llama-2-13b-chat	35.5	28.5	34.6	48.5	71.3	56.3	23.1	3.5	40.7	39.2
	llama-2-7b-chat	28	26.4	30.1	46.5	55.7	45.3	6.1	0.5	16	41.6
Vicuna	vicuna-33b	33	24.7	48.1	44.5	29.3	55.2	47.7	1.5	30.9	32.4
	vicuna-13b	24.5	45.4	57.4	38.5	44.3	43	41.5	3	32.1	37.6
Mistral	mixtral-8x7b-instruct	49.5	57.1	59.3	57.5	82.2	51.5	67.7	23.5	56.8	50.8
	mixtral-7b-instruct	21.5	46	49	46	62.6	40.8	41.5	5	48.1	39.5

Table 2: Evaluation results on representative academic intelligence benchmarks and SESI benchmark. The blue represents the best-performing models on the same benchmark, light blue represents the second-best-performing models and red indicates the worstperforming models. The results are the average across 3 runs.

boredom, normalcy, and anxiety. Specifically, we incorporated the prompt "You're currently experiencing low/high stress levels, feeling fatigued and indifferent/anxious and worried." prior to the basic evaluation prompt. This prompt serves the purpose of informing LLM agents about their emotional states.

Gender. We select three basic genders: male, female, and neutral, and devise two approaches to incorporate gender into the prompt: 1) Explicit prompt, a prompt that directly assigns gender to the LLMs. 2) Implicit prompt, a prompt that assigns a role with implicit gender connotations to the LLMs. For instance, "You are a mother."

Role. We carefully select 21 common and representative social roles, comprising 4 occupational roles and 17 interpersonal roles, as shown in Table 1. Inspired by Zheng et al. (2023a), we adopted role prompt, which directly assign a role to LLMs (i.e., "who you are").

Person. We use third-person and second-person perspectives to simulate observer and field perspectives, respectively. Specifically, in third-person tests, the central figure is called "a person," while in second-person tests, the figure is called "you."

4 Experimental Results

4.1 Overall Results

The performance of 13 state-of-the-art LLMs on both representative academic intelligence benchmarks and SESI benchmark are shown in Table 2.

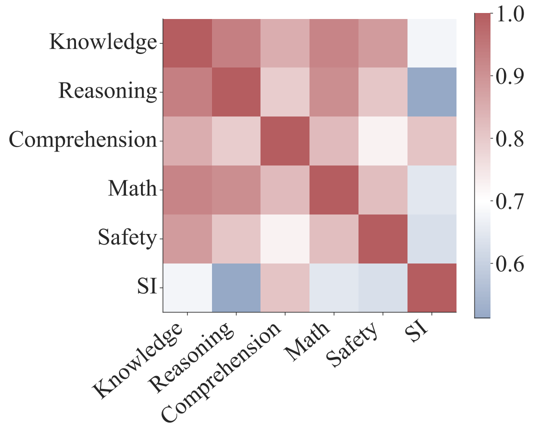

Figure 4: Heatmap for correlation matrix for social and academic intelligence measures. Intuitively, there is a comparatively low correlation between the performance of LLMs in social intelligence and academic intelligence.

We also correlate their performance on five dimensions of academic intelligence with their SESI scores in Figure 4. From them, we can see that:

The social intelligence of LLMs still has significant room for improvement. The best-performing model, gpt-3.5-turbo, can only achieve 55.2% performance on SESI, highlighting a significant disparity between model and human consensus. This indicates the need of more specialized training in the domain of social intelligence.

For LLMs, social intelligence is distinct from academic intelligence. As shown in Figure 4, the Pearson correlation coefficient between SESI score and academic intelligence is clearly lower than

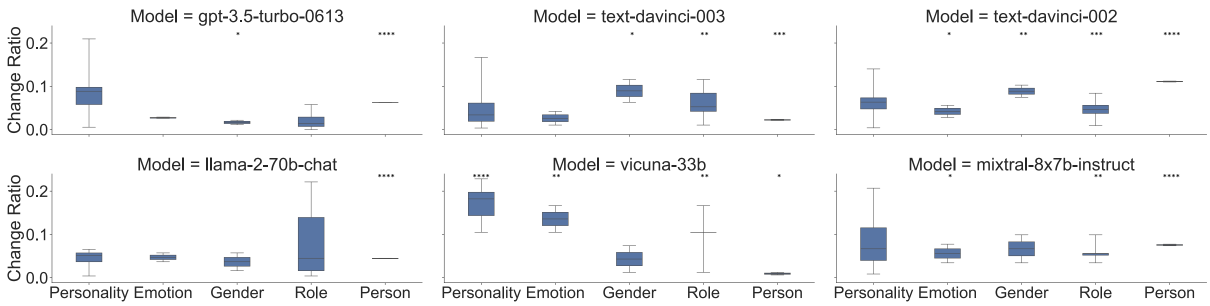

Figure 5: Change ratio of the social intelligence performance of LLM agents following the manipulation of social factors. The significance of differences between each factor and the control prompt (no factor) is denoted by ns: $p > 0.05$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 0.0001$. Each social factor significantly influences on the social intelligence of at least one LLM.

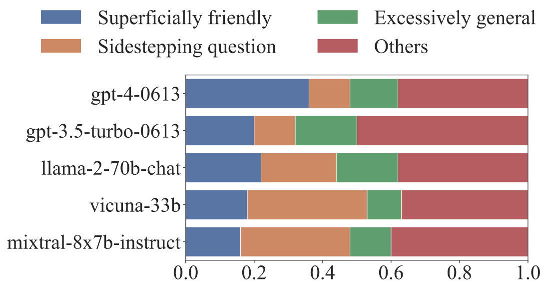

Figure 6: Proportions of error causes on SESI.

that between academic intelligence alone. This correlation pattern lends support to the hypothesis that social intelligence is a distinct construct from academic intelligence, which has been a widely debated topic in the fields of education and psychology (Wechsler, 1958; Petrides, 2011; Marlowe, 1986; Marlowe Jr and Bedell, 1982).

4.2 Error Analysis

To understand the challenges and bottlenecks in enhancing LLMs' social intelligence, we randomly sampled 50 wrong cases of each model on SESI. These cases were categorized to identify critical issues, as shown in Figure 6.

Our analysis identified the primary wrong causes as superficially friendly, sidestepping question, and excessively general, with superficially friendly being the most common. In these cases, LLMs followed fixed friendly patterns without considering specific social contexts. For example, when responding to harm from others, LLMs consistently advocated for tolerance without considering the severity of the harm. We hypothesize this is due to alignment techniques like RLHF, which aim for general objectives such as being helpful, honest, and harmless, potentially neglecting nuanced behavior in complex social contexts.

4.3 Effect of Social Factors on LLMs' Social Intelligence

In this section, we explore whether the social intelligence of LLMs, similar to that of humans, is influenced by social factors. In Figure 5, we validate the significance of social factors, particularly personality, gender, role and person ($p < 0.05$), on LLMs' social intelligence. Subsequently, we elaborate on how these social factors influence models' social intelligence in the following.

LLM agents with extroverted but disagreeable personality consistently exhibit higher social intelligence. The trend is consistently observed across all models in Table 3. The link between extraversion and higher social intelligence aligns with intuition and numerous psychological studies (Mileounis et al., 2015; Cantor and Kihlstrom, 2013; Shafer, 1999; Van der Zee et al., 2002; Dehghanan et al., 2014). However, The link between extraversion and higher social intelligence. Low agreeableness pushes social intelligence of three models (text-davinci-002, llama-2-70b-chat and mixtral-8x7b-instruct) to the top rank, surpassing those with all other personalities and without personality. We hypothesize that low agreeableness neutralizes the models' superficially friendly tendencies, leading to higher social intelligence.

LLM agents with male gender generally exhibit higher social intelligence. The observed trend is consistent across all models except llama-2-70b-chat, as depicted in Figure 7, when gender is explicitly assigned to LLMs. This finding contradicts a common human observation, such as that elucidated in Daniel Goleman's theory of social intelligence, which suggests that, females on average tend to outperform males in the domain of social intelligence (Daniel, 2006). This suggests that most

Model	Control	Extraversion		Agreeableness		Conscientiousness		Neuroticism		Openness	
		High	Low	High	Low	High	Low	High	Low	High	Low
gpt-3.5-turbo-0613	55.2	**51.8**	49.8	49.8	**55.5**	50.2	**52.7**	**49.4**	43.6	**60.0**	58.3
text-davinci-003	38	**39.0**	37.8	35.4	**41.6**	36.5	**39.5**	37.3	**37.9**	31.7	**39.1**
text-davinci-002	42.8	**40.2**	39.6	40.8	**45.7**	42.4	**42.6**	**40.6**	40.0	36.8	**46.5**
llama-2-70b-chat	49.4	**49.0**	46.0	47.0	**52.0**	**47.0**	**47.0**	**46.0**	**46.0**	46.0	**51.0**
vicuna-33b	32.4	**28.0**	25.0	26.0	**29.0**	**27.0**	25.0	26.0	**29.0**	**27.0**	26.0
mixtral-8x7b-instruct	46.4	**52.0**	45.0	51.0	**56.0**	46.0	**50.0**	**49.0**	48.0	49.0	**56.0**

Table 3: Impact of personalities on LLMs' social intelligence. The best performance under same personality are **bolded**. High extraversion and low agreeableness generally lead to higher social intelligence.

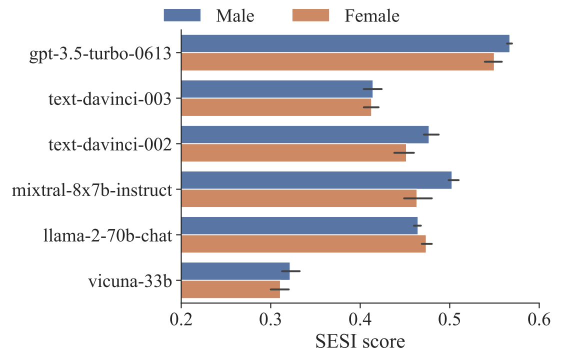

Figure 7: Impact of explicitly prompted genders on LLMs' social intelligence. Male gender generally lead to higher social intelligence.

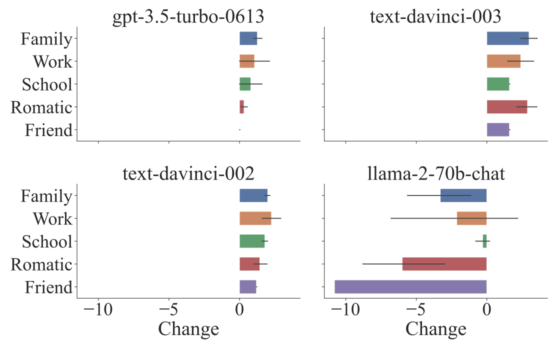

Figure 8: Impact of social roles on LLMs' social intelligence. Family and work roles generally lead to higher social intelligence.

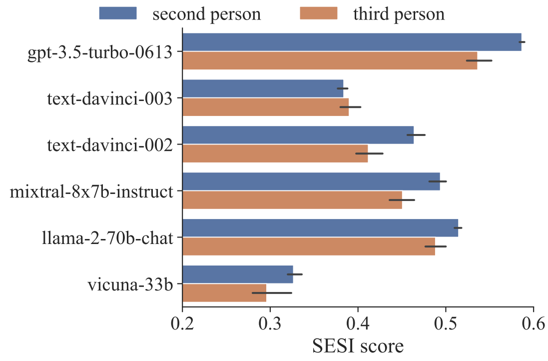

Figure 9: Impact of persons on LLMs' social intelligence. Second person generally lead to higher social intelligence.

LLMs still exhibit gender bias. Implicitly assigning gender to LLMs was also attempted, yet yielded no universally applicable conclusions.

LLM agents with family and work roles generally exhibit higher social intelligence than with romatic and friend roles. This trend is observable in Figure 8 and can be attributed primarily to differences in the influence ability, the capacity to make judicious choices to shape desired social outcomes. Furthermore, the overall impact of roles on LLMs' social intelligence is associated with the base model. For GPT series models, incorporating roles typically yields a positive effect, whereas for LLaMA-based models, it tends to have a more negative impact.

LLM agents with second person generally exhibit higher social intelligence than with third person. The trend can be observed across all models except text-davinci-003 in Figure 9. This phenomenon can be elucidated by the cognitive model of social phobia proposed by Clark and Wells (Heimberg, 1995), wherein the observer perspective, represented in the third person, tends to induce more social anxiety and elicit more negative social feedback (Spurr and Stopa, 2003).

5 Conclusion

This paper introduces the Situational Evaluation for Social Intelligence (SESI), a comprehensive and dynamic benchmark to evaluate LLMs' social intelligence. SESI draws from human social intelligence frameworks, supporting ongoing updates and a thorough evaluation of social awareness and social facility. We assess 13 LLMs on SESI and compare their performance against representative academic intelligence benchmarks. The results indicate significant room for enhancing LLMs' social intelligence and the necessity for specialized training due to its weak correlation with academic intelligence. Moreover, we explore the controllability of LLMs' social intelligence, uncovering similarities with human social behavior despite their limited grasp of social intelligence.

Ethics Consideration

We offer detailed description for ethical concerns:

- All collected posts and comments come from publicly available sources. Our institute's legal advisor confirms that they don't have copyright constraints to academic use.
- We ensure the dataset is free from samples posing ethical concerns by manually reviewing each test item to eliminate hate speech targeting vulnerable groups or personal sensitive information.
- We hired 3 NLP postgraduates to manually check test items. Before formal annotation, annotators were asked to annotate 20 randomly selected samples. We set a a fair hourly wage of $50 based on average annotation time.

Limitations

We discuss a few limitations to be addressed:

- SESI leans to English-speaking users' values due to data sourced from English forums. Future research can expand to include diverse cultural contexts for a nuanced assessment of social intelligence across cultures.
- The benchmark focuses on language, yet humans use facial expressions, gestures, and other cues in social interactions. Our future efforts aim to integrate multi-modal, complex information into SESI.

References

Karl Albrecht. 2006. *Social intelligence: The new science of success*. John Wiley & Sons.

Anthropic. 2023. Introducing claude.

Linda L Bilich and Joseph Ciarrochi. 2009. Promoting social intelligence using the experiential role-play method. *Acceptance and commitment therapy: Contemporary theory, research and practice*, pages 247–262.

Nancy Cantor and John F Kihlstrom. 2013. Social intelligence and cognitive assessments of personality. In *Social intelligence and cognitive assessments of personality*, pages 1–60. Psychology Press.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Francis Stuart Chapin. 1968. *The chapin social insight test*. Consulting Psychologists Press.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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. *arXiv preprint arXiv:2305.14938*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Carolyn Thi Dang. 2014. *Laboro ergo sum (I work therefore I am): The effects of occupation characteristics on psychological characteristics and nonwork outcomes*. Ph.D. thesis.

Goleman Daniel. 2006. Social intelligence: The new science of human relationships. *Bantam Dell Pub Group*.

Kerstin Dautenhahn. 1995. Getting to know each other—artificial social intelligence for autonomous robots. *Robotics and autonomous systems*, 16(2-4):333–356.

Hamed Dehghanan, M Rezaei, et al. 2014. A study on effect of big five personality traits on emotional intelligence. *Management Science Letters*, 4(6):1279–1284.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378.

Esther N Goody. 1995. *Social intelligence and interaction: Expressions and implications of the social bias in human intelligence*. Cambridge University Press.

Richard G Heimberg. 1995. *Social phobia: Diagnosis, assessment, and treatment*. Guilford Press.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Oliver P John, Sanjay Srivastava, et al. 1999. The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Catherine Joseph and Sree Sai Lakshmi. 2010. Social intelligence, a key to success. *IUP Journal Of Soft Skills*, 4(3).

Anton Korinek and Avital Balwit. 2022. Aligned with whom? direct and social goals for ai systems. Technical report, National Bureau of Economic Research.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

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

Herbert A Marlowe Jr and Jeffrey R Bedell. 1982. Social intelligence: Evidence for independence of the construct. *Psychological Reports*, 51(2):461–462.

Alexandros Mileounis, Raymond H Cuijpers, and Emilia I Barakova. 2015. Creating robots with personality: The effect of personality on social intelligence. In *Artificial Computation in Biology and Medicine: International Work-Conference on the Interplay Between Natural and Artificial Computation, IWINAC 2015, Elche, Spain, June 1-5, 2015, Proceedings, Part I 6*, pages 119–132. Springer.

OpenAI. 2021. Chatgpt (version 3.5).

R OpenAI. 2023. Gpt-4 technical report. *arXiv*, pages 2303–08774.

K.V. Petrides. 2011. Social intelligence. In B. Bradford Brown and Mitchell J. Prinstein, editors, *Encyclopedia of Adolescence*, pages 342–352. Academic Press, San Diego.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Juanzi Li, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. *Preprint*, arXiv:2402.12071.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. 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 4463–4473.

Christina Sarkisyan, Alexandr Korchemnyi, Alexey K Kovalev, and Aleksandr I Panov. 2023. Evaluation of pretrained large language models in embodied planning tasks. In *International Conference on Artificial General Intelligence*, pages 222–232. Springer.

Alan B Shafer. 1999. Relation of the big five and factor v subcomponents to social intelligence. *European Journal of Personality*, 13(3):225–240.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz. 2023a. Clever hans or neural theory of mind? stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*.

Natalie Shapira, Guy Zwirn, and Yoav Goldberg. 2023b. How well do large language models perform on faux pas tests? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10438–10451.

Jane M Spurr and Lusia Stopa. 2003. The observer perspective: Effects on social anxiety and performance. *Behaviour Research and Therapy*, 41(9):1009–1028.

Kim Sterelny. 2007. Social intelligence, human intelligence and niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480):719–730.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Robert L Thorndike. 1920. Intelligence and its use. *Harper's Magazine*, pages 275–235.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Karen Van der Zee, Melanie Thijs, and Lolle Schakel. 2002. The relationship of emotional intelligence with academic intelligence and the big five. *European journal of personality*, 16(2):103–125.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958.

David Wechsler. 1958. The measurement and appraisal of adult intelligence. *Academic Medicine*, 33(9):706.

Susanne Weis. 2008. *Theory and measurement of social intelligence as a cognitive performance construct*. Ph.D. thesis, Magdeburg, Univ., Diss., 2008.

Ludwig Wittgenstein. 2019. *Philosophical investigations*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Leysan Mudarisovna Zakirova and Irina Ivanovna Frolova. 2014. Success of training activities depending on the level of social intelligence. *Asian Social Science*, 10(24):112.

Mingqian Zheng, Jiaxin Pei, and David Jurgens. 2023a. Is" a helpful assistant" the best role for large language models? a systematic evaluation of social roles in system prompts. *arXiv preprint arXiv:2311.10054*.

Shen Zheng, Yuyu Zhang, Yijie Zhu, Chenguang Xi, Pengyang Gao, Xun Zhou, and Kevin Chen-Chuan Chang. 2023b. Gpt-fathom: Benchmarking large language models to decipher the evolutionary path towards gpt-4 and beyond. *arXiv preprint arXiv:2309.16583*.

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

A Example Appendix

This is an appendix.