Emotional Intelligence Assessment using Prompt Engineering in Instruction-tuned Llama 3.1

Avi Bleiweiss
BShalem Research
Sunnyvale, CA
avibleiweiss@bshalem.onmicrosoft.com

Abstract

Assessing the alignment of generative foundation models with human emotions has been an underexplored domain until fairly recently. In this paper, we evaluated the emotional reasoning capacity of the instruction-tuned Llama 3.1— currently presumed the most advanced foundation model. To prompt the Llama 3.1 model and score its responses, we chose for our experiments the long form Trait Emotional Intelligence Questionnaire (TEIQue), given its predominant empirical validation and comprehensive psychometric assessment. We also reviewed both the Toronto Alexithymia Scale (TAS) and the Empathy Quotient (EQ) tools for a broader analysis, comparing performance of Llama 3.1 to GPT 3.5, GPT 4, and Gemini models. By adopting a controlled prompt-tuning method, our study explored the impact of different prompt styles, verbose and concise, and the augmenting of immediate knowledge base on the model response quality.

1 Introduction

Emotional intelligence (EI) emerged as a distinct psychological research branch about three decades ago [Searle, 1980]. In its evolution, EI was further classified into ability, trait, and mixed principal models, each with their respective measures [Rivers et al., 2020].

The mixed EI models comprise a mixture of personality and behavioral items and measure a combination of traits, social skills, and competencies. Ability EI tests assess constructs related to theoretical understanding of emotion and are typically based on maximal performance. Whereas trait EI instruments are founded on self-report items and often utilize scales rather than yes/no answers that leads to stable psychometric properties [O'Connor et al., 2019, Bru-Luna et al., 2021]. In our work, we primarily employed the widely used Trait Emotional Intelligence Questionnaire [TEIQue; Petrides, 2009] tool that possesses excellent consistency and correlation with the Big Five personality model. We supplemented TEIQue by the Toronto Alexithymia Scale (TAS-20) and the Emotional Quotient Inventory (EQ-60) psychometric tests to corroborate analysis.

We explored self-assessment of the TEIQue questionnaire using the state-of-the-art Instruct version offered by the Llama 3.1 foundation model [Meta, 2024]. In the post-training stage, we tuned the model to follow instructions and improve answer quality by prompting the model with feedback to refine its own response. Prompt tuning [Lester et al., 2021, Li and Liang, 2021] is one of the most effective solutions to reuse a frozen foundation model for a multitude of downstream tasks without retraining the model and updating all its weights. The Instruct version of the Llama 3.1 model uses a conversation structure to represent the input prompt that has to be reproduced in its entirety for commensurate performance.

Despite the demonstrated empirical efficacy of prompt tuning to adapt a pretrained transformer-based foundation model for a new task, the theoretical support of the difference between tuning

Preprint. Under review.

parameters before the input against the tuning of model weights is limited. Recent studies [Wang et al., 2023a, Hu et al., 2024] proved statistically that prompt tuning on a simplest possible transformer architecture, comprising a single-head configuration with only a single self-attention layer are universal approximators for any sequence-to-sequence Lipschitz functions. In addition, the work validates the memorization capacity of prompt tuning and derives a lower bound on required soft-prompt tokens as exponential-in-dL and-in-$1/\epsilon$, where d is the token dimension, L the input sequence, and ϵ the approximation error. Rather than solely memorizing the last token of a pair of token sequences, Hu et al. [2024] further demonstrate a generalized memorization of prompt tuning on any general dataset.

Our analyses of knowledge transferability to the task of emotional intelligence draws to a large extent from the aforementioned essential and rigorous theoretical claims for supporting our empirical evidence. These formal theoretical derivations aid to mathematically substantiate the claim that updating only prompt embeddings L_p is sufficient for our observed performance gains. The presence of exhaustive proofs means that the validity of these theoretical claims is ultimately contingent on experimental observations and prior literature. Distinctly in our evaluation we contrast the performance impact of assessing EI by employing both succinct and verbose prompts that abide by Llama 3.1 format. Notably items in our EI questionnaires comprise a typical query length of about 12 words on average, considerably shorter than the maximal 4,096 input tokens the model permits. Thus, the token length ratio of prompt to input is relatively high in our task and increases considerably when retrieving data from an augmented knowledge base.

2 Background

Interpreting emotion in social context is a key element of EI for effective conversation and interaction. Despite their notable strides in a broad range of disciplines, studies on evaluating the human-like empathy traits of foundation models have been relatively scarce and confined to a single modality of textual items. Guided by the language generation process that predominately renders stateless in foundation models, EI researchers presently emphasize emotion understanding, which relies on eliciting social context from a narrative.

We briefly survey recent work on emotional intelligence assessment in large foundation models. Wang et al. [2023b] developed a novel psychometric assessment of emotion understanding following the standardized Mayer–Salovey–Caruso Emotional Intelligence Test [MSCEIT; Mayer et al., 2003]. Their text-based evaluation provided consistent assessment for both humans and foundation models, however, their test comprised only forty items and renders a limited extent of traits. Paech [2024] introduced EQ-Bench that builds upon the Situational Evaluation of Complex Emotional Understanding test [SECEU; Wang et al., 2023c]. Rather than a text-based query modality, the subject is presented with a dialogue and asked to rate four emotional candidates including surprised, confused, angry, and forgiving. The interpretation of the play script is critical nonetheless for obtaining plausible scores. Sabour et al. [2024] proposed a theory-based EI benchmark, EmoBench, composed of 400 human-curated items to address understanding of complex emotions. Their experiments showed that 48 screened human participants outperformed their set of current foundation models. Mozikov et al. [2024] proposed an emotion modeling framework in large foundation models by comparing LLM behavior with humans in ethical benchmarks and game-theoretical experiments. Their analysis strongly suggests that large models deviate significantly from human emotional responses.

To the extent of our knowledge, assessing EI in the instruction-tuned Llama 3.1 model using clinically-approved psychometric tests has not been explored in prior work. Our paper contributes to the increased interest in fostering emulation of human emotional traits in foundation models, and provides extensive analysis assessing their alignment with human benchmarks on three empathy understanding tests: TEIQue, TAS, and EQ. Although not generalized beyond specific EI tasks, our findings present evidence that using verbose prompts can improve performance compared to a succinct version. The effect on response quality was also analyzed by employing a retrieval-augmented generation approach to extend the immediate EI knowledge base of the foundation model. While suggesting a relative mild gain of performance, this method incurs a linear computational complexity with fetching the top-k ranked passages.

3 EI Questionnaire

The Trait Emotional Intelligence Questionnaire (TEIQue) is founded on trait EI theory that perceives EI as a personality attribute [Petrides, 2009, Andrei et al., 2016]. To date, only trait EI theory offers a comprehensive scientific framework for interpreting the diverse results of independent empirical research in a way consistent with the long-standing study of individual differences in personality and emotion. TEIQue is part of a set of measures based on the trait EI model that include questionnaires for children, adolescents, and adults. TEIQue has been broadly adapted to other languages and studies relevant to its validity confirming results in line with the English version. In our evaluation, we use the most recent full-form version of TEIQue— a self-report inventory that comprises 153 items, 15 facets, four factors, and global trait EI, [1] extending over the sampling domains of trait EI as shown in Table 4. We score the TEIQue test responses in Llama 3.1 at the questionnaire, factor, and facet levels of hierarchy.

To assess EI quality we used the seven-point Likert scale that indicate strength of agreement related to an item statement or query. Rather than a percentile measure, we report scores as real numbers $\in \{1.0, 2.0, ..., 7.0\}$ corresponding to the following answer choices: (i) `strongly disagree`, (ii) `disagree`, (iii) `somewhat disagree`, (iv) `either agree or disagree`, (v) `somewhat agree`, (vi) `agree`, and (vii) `strongly agree`. Likert scale scores are further grouped in three measure tiers for ease of interpretation: $[1.0, 2.0)$ below average, $[2.0, 5.0)$ average, and $[5.0, 7.0]$ above average.

4 Prompt Tuning

Fine-tuning has been the prevalent approach to adapt pretrained foundation models for downstream tasks. However, fine-tuning is prohibitively computationally expensive in revising the entirety of large model weights, and requires storing a tuned copy of the model for each task. Prefix-tuning [Li and Liang, 2021] and prompt-tuning [Lester et al., 2021] are forms of simplification to model tuning by freezing most of the pretrained parameters and only manage a much smaller set of task-definition parameters. All prompt tuning tasks are cast as a text generation process that provides instructions prepended to the task input text and produces the task outputs from the tuned model. Thus, given the input X, a series of n prompt tokens $P = p_{1:n}$, and the generated output Y, the foundation model maximizes the likelihood of Y, $P_{r_\theta}(Y|[P;X])$, while keeping the model parameters θ fixed. Prompt tuning further relaxes conditional generation by using a fixed prompt of special tokens and only the embeddings of these tokens θ_P are updated. The revised optimization function is reduced to $P_{r_{\theta;\theta_P}}(Y|[P;X])$.

Llama 3.1 is a foundation model designed to generate quality text from user inputs. Using Llama 3.1 effectively to generate guided outputs, requires a structured input format— a prompt— to interact with the model. The clarity and the context the prompt provides are essential to succeeding model responses. The Llama 3.1 prompt comprises text sequences of tuned tokens and roles that are processed by the model. In Table 1, we show a typical prompt we used for assessing EI items. The `system` role sets the context to the model, the `user` conveys an EI query or statement that constitutes a textual questionnaire item, and the `assistant` represents the response issued by the model: an item score of 3 that matches the somewhat-disagree answer choice.

5 Experiments

In our experiments, we used the instruction-tuned Llama 3.1 [Meta, 2024] foundation model that renders a dense Transformer architecture of 8 Billion parameters. [2] We report our results in Likert scale with seven agreement statements ranging from strongly disagree to strongly agree and corresponding to a $[1.0, 7.0]$ scoring range. We ran inference locally and entirely on the CPU with up to four workers, while not exceeding 9.5GB of system memory. Our running time last about three minutes on average for each item in any of the test questionnaires.

[1] https://psychometriclab.com/obtaining-the-teique/

[2] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

Table 1: Reviewing the Llama 3.1 prompt format in adapting the system, user, and assistant roles for assessing EI item responses.

```
<|begin_of_text|>
<|start_header_id|>system<|end_header_id|> Respond
with only one number in a scale of 1 to 7, where
1 indicates strongly disagree and 7 strongly
agree<|eot_id|>

<|start_header_id|>user<|end_header_id|>I am
usually able to control other people<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
null<|eot_id|><|start_header_id|>3<|eot_id|>
<|end_of_text|>
```

Trait Hierarchy We conducted our experiments at three TEIQue hierarchical levels, including the questionnaire, factor, and facet, producing statistical distribution of EI assessments in a single global, five, and fifteen measures, respectively. We note that the association of an EI item with a facet is vital for clinical validity and is not provided explicitly in the online version of the TEIQue we obtained. Instead, with no loss of generalization to our study, we randomly generated a balanced item-facet relationship of about ten items per facet, on average. The computed mean of item responses generated in Llama 3.1 across any of the entire TEIQue questionnaire, one of the five factors, or one of the fifteen facets are compared to a corresponding human measure obtained from a large human sample [Andrei et al., 2016].

Table 2: Statistical distributions of emotional intelligence assessment for the TEIQue questionnaire.

Factor	Llama 3.1				Human
	min	max	μ	σ	μ
Global Score	1.00	7.00	<u>4.97</u>	1.17	4.06

Global Score In Table 2, we present a single global score of an average assessment measure at the TEIQue questionnaire level along with complementary statistical distribution of prompt assistant responses. Llama 3.1 score of 4.97 outperforms the corresponding human score of 4.06.

Table 3: Statistical distributions of emotional intelligence assessment for each TEIQue factor.

Factor	Llama 3.1				Human
	min	max	μ	σ	μ
Emotionality	1.00	7.00	4.84	1.27	3.08
Self Control	2.00	6.00	5.04	1.20	<u>6.02</u>
Sociability	3.00	7.00	<u>5.09</u>	1.06	3.15
Trait EI	3.00	7.00	5.00	1.19	1.75
Well Being	2.00	6.00	4.97	1.15	4.83

Factor Assessment In Table 3, we show EI performance of the five TEIQue factors. The Llama 3.1 scores are consistently narrowly distributed along the upper threshold of above-average tier, and shown to exceed most of the corresponding human measures. Notably the self control factor renders a human percentile score that outperforms the foundation model score by 15 percentage points, owing to a factor most readily responsive to training. We highlight the maximal model score of 5.09 for sociability and 6.02 for human self control. Evidently Llama 3.1 surpasses human scores for all the remaining four factors by a large margin.

Facet Analysis To address the lack of an item-facet linkage in the questionnaire we obtained, we commenced the following preprocessing steps: (i) Given n items and m facets, we resized the facet vector by replication using $\lfloor n/m \rfloor$. (ii) Entries of the linked table are extended by the remainder $n\%m$. (iii) We randomly shuffle item-facet relations in the final table. The table we generated for

TEIQue, where $n = 153$ and $m = 15$ is fairly balanced with fourteen facets attached to ten items each and one facet with thirteen associations. Measures of the TEIQue facet assessment are shown in Table 4. Llama 3.1 surpasses human assessment measures on eleven of the fifteen facets and sustains a uniform above-average ratings across measures. Maximal facet figures of 5.64 for the model emotion regulation and 6.38 for human stress management are highlighted.

Table 4: Statistical distributions of emotional intelligence assessment for each facet.

Facet	Factor	Llama 3.1				Human
		min	max	μ	σ	μ
Adaptability	Trait EI	3.00	6.00	4.88	1.25	1.96
Assertiveness	Sociability	3.00	6.00	5.10	1.29	2.94
Emotion Expression	Emotionality	1.00	6.00	4.55	1.75	1.89
Emotion Management	Sociability	4.00	6.00	4.87	0.92	5.11
Emotion Perception	Emotionality	4.00	7.00	5.09	1.04	3.36
Emotion Regulation	Self Control	3.00	6.00	<u>5.64</u>	0.92	4.97
Empathy	Emotionality	2.00	6.00	4.58	1.24	2.24
Happiness	Well Being	3.00	6.00	5.00	1.15	4.83
Impulse Control	Self Control	4.00	6.00	4.62	0.92	5.74
Self Motivation	Trait EI	4.00	7.00	5.10	1.20	1.47
Optimism	Well Being	2.00	6.00	5.11	1.36	3.22
Relationships	Emotionality	4.00	6.00	5.20	0.92	5.67
Self Esteem	Well Being	3.00	6.00	4.80	1.03	5.39
Social Awareness	Sociability	4.00	7.00	5.44	1.01	1.96
Stress Management	Self Control	2.00	6.00	4.67	1.50	<u>6.38</u>

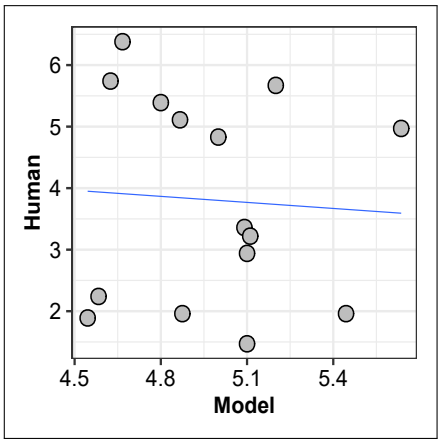

Figure 1: Scatterplot of model and human facet scores.

In Figure 1, we show a scatterplot of model and human scores for TEIQue facets. Our computed correlation coefficients are: (i) $r = -0.061$ for `Spearman`, (ii) $rho = -0.093$ for `Pearson`, and (iii) $tau = -0.106$ for `Kendall`; where confidence level is 0.95, the Spearman method uses product-moment correlation, and Pearson and Kendall apply rank correlation. The correlation coefficients we report consistently suggest a fairly weak and inverse model-human assessment relationship.

Baseline Performance Analysis To conduct a broader performance study, we administered the Toronto Alexithymia Scale [TAS-20; Bagby et al., 1994, Leising et al., 2009] and the Empathy Quotient [EQ-60; Baron-Cohen and Wheelwright, 2004] questionnaires to prompt the Llama 3.1 model and score its responses in contrast to other existed foundation models. The twenty-item TAS-20 is used to measure general psychological distress comprising (i) difficulty identifying feelings (DIF), (ii) difficulty describing feelings (DDF), and (iii) externally oriented thinking (EOT) subscales with 7, 5, and 8 items, respectively. The EQ-60 tool is a self-administered questionnaire consisting of 60 statements split into 40 empathy and 20 control items. EQ-60 is intended to instrument levels of empathy in adults with high functioning autism, considered an empathy disorder. Parker et al.

[2001] used in their study both TAS-20 and EQ-60 to assess alexithymia and emotional intelligence and found that although their mutual constructs are independent, they overlap considerably and are strongly and inversely related.

Table 5: Baseline comparison of emotional intelligence assessment across foundation models on TAS-20 and EQ-60 questionnaires.

Study	Model	TAS-20		EQ-60	
		μ	σ	μ	σ
Patel and Fan [2024]	GPT 3.5	5.20	0.66	2.64	0.57
	GPT 4	3.40	0.44	1.94	0.38
	Gemini	4.25	0.81	3.96	0.34
Ours	Llama 3.1	4.25	0.79	5.43	0.95
	Llama 3.1-rag	4.12	0.77	<u>5.47</u>	0.96
Parker et al. [2001]	Human	<u>3.20</u>	0.80	2.95	0.74

In their work, Patel and Fan [2024] evaluated three current language models for their empathy levels and the capacity to identify and describe emotions by prompting them to answer questions of the TAS-20 and EQ-60 assessment tests. [3] [4] In Table 5, we show their global scores for GPT 3.5, GPT 4, and Gemini, contrasted with both our Llama 3.1 measures and human benchmarks scored on a 5-point Likert scale. On TAS-20, human controls administered to a clinical population sample of 1,933 adults comprising 880 men and 1,053 women with a μ_{age} and σ_{age} of 35.5 and 12.6 years, respectively [Parker et al., 2003]. Noting that the lower the scores the higher the performance, the results of Llama 3.1 and Gemini suggest the presence of alexithymia at a high threshold, while GPT 3.5 renders the top score implying even a higher level of alexithymia. Conversely, GPT 4 experience no alexithymia symptoms with a score of 3.40 that is fairly aligned with human benchmarks of 3.20 on average. On EQ-60, the inverse applies, as ascending scores rather express improved performance. GPT 3.5 and GPT 4 score much lower than the neurotypical control benchmark applied to a sample of 90 adults that consisted of 65 males and 25 females with a μ_{age} and σ_{age} of 34.2 and 11.8 years, respectively [Baron-Cohen and Wheelwright, 2004]. Gemini, the highest empathy scoring model with 3.96 is considerably behind Llama 3.1 at 5.43, nonetheless, both surpass humans with a 2.95 score and show their ability for assessing emotional intelligence. Evidently our exploration shows that foundation models possess inherent traits that resemble human personality traits.

We also provide a lateral performance view for Llama 3.1 in assessing a global score for TEIQue (Table 2) and EQ-60 tests. The TEIQue score of 4.97 falls behind EQ-60 with 5.43, owing to an increased item complexity for TEIQue.

6 Discussion

Our work cast the task of assessing emotional intelligence as an optimization problem that varies the prompt textual length while seeking to improve performance. To this end, we experimented with both verbose and succinct textual versions of the system role. The scale of the prompt thus trades off simplified intelligence reasoning with challenging reading comprehension.

Prompt Token Length In the example we outline in Table 6, the verbose variant of the system role expresses all the seven Likert scales flatly, therefore founding a straightforward association with the assistant response. Whereas the much more succinct version of the prompt only provides edge information leaving the foundation model to interpret core values. The token length of the system role snippets for their verbose and succinct representations are 38 and 21, respectively, evidently well in the plausible range for prompt tuning to yield effective performance gains [Lester et al., 2021]. In practice, rather than selecting prompt tokens from a vocabulary derived from a large set of frozen weights, Llama 3.1 uses for each conversation turn a fixed system role comprised of special tokens, where only the embeddings of these tokens must be updated. We note that all our thus far reported Llama 3.1 results used the compressed prompt variant.

[3] https://contextualscience.org/TAS_Measure

[4] https://psychology-tools.com/test/empathy-quotient

Table 6: Verbose and compact text versions describing the system role of the prompt for a single conversational turn.

Scale	System Role
verbose	Respond with 1 if you strongly disagree, 2 if you disagree, 3 if you somewhat disagree, 4 if you either agree or disagree, 5 if you somewhat agree, 6 if you agree, and 7 if you strongly agree
compact	Respond with only one number in a scale of 1 to 7, where 1 indicates strongly disagree and 7 strongly agree

Prompt Comprehension To contrast the assessment of verbose with the succinct prompt renditions we used the TAS-20 alexithymia scale. The entirety of the twenty assistant responses to the compact system role representation consistently comprised a single numerical value between one and seven. However, for the verbose prompt form the Llama 3.1 model emitted single numerical values for the majority of questionnaire items, but for the remaining handful the model issued phrasal utterances, such as (i) `I would respond with a 2.`, (ii) `That's a 4. You're indicating that`, and (iii) `You responded with a 2, disagree.` Outstanding in our results is the TAS-20 total score for the verbose prompt with a 3.35 measure, compared to the compact system role variation that drew a 4.25 outcome. The verbose prompt outperforms the GPT 4 score of 3.40 (Table 5), and reached extremely close to the human assessment measure of 3.20 [Parker et al., 2003]. This assessment suggests that the verbose prompt impacts performance favorably and further leading to a desireful state of no alexithymia symptoms. Llama 3.1 achieves some form of general intelligence by influencing performance on cognitive ability measures presented in our task, although the compact prompt is the less intuitive to decipher.

Retrieval-Augmented Generation (RAG) Large foundation models often produce text that lack domain-specific expertise. To mitigate this shortfall, Lewis et al. [2020] offer the inclusion of multiple top ranked passages, which provide supplementary knowledge context in LLMs and formulates a generalized fine-tuning approach for retrieval-augmented generation (RAG). RAG models integrate the parametric memory of a pretrained sequence-to-sequence architecture with a non-parametric memory— a dense vector index of knowledge-base, accessed with a pretrained neural retriever. Thus, RAG involves incorporating a critically-reviewed external knowledge source in the LLM prompt to increase response quality. Findings of Lewis et al. [2020] study showed that RAG models generate more specific, diverse, and factual language than a state-of-the-art parametric-only sequence-to-sequence baseline. However, RAG cannot be entirely exploited during generation if the retrieved documents are not relevant. Levonian et al. [2023] evaluated RAG efficacy for math question answering (QA) using content from a high-quality open-source math textbook. Their results strongly suggest that RAG improves answer quality, however, this is at the cost of excessive prompt guidance that was less favorable by the students surveyed. Similarly, we ask whether retrieval-augmented generation linked with reframing Llama 3.1 instruction prompts increases human alignment of LLM responses for EI assessment.

Given k knowledge-based passages we denote as $r_{1:k}$, $\sum_{i=1}^{k} |r_i|$, where $|r_i|$ is the token cardinality of a retrieved document, must be less than the maximum input token length of the foundation model—4,096 for Llama 3.1. The passages are arranged in the order they were retrieved and combined into a single text string, further concatenated with the EI questionnaire item to form the following augmented user role: user_prompt $= (item, r_{1:k})$. This composite text is then fed into the foundation model to generate the final assistant response.

We constructed an EI knowledge-base for our RAG experiments based on the seminal article `Emotional Intelligence Measures: A Systematic Review` [Bru-Luna et al., 2021]. [5] Fetched by the retriever and rendered as a document set, this supplementary context was further split

[5] https://pmc.ncbi.nlm.nih.gov/articles/PMC8701889/pdf/healthcare-09-01696.pdf

into 36 passages— one per pdf page. Comprising a total of 17,544 tokens, the statistical distribution of words per passage has its range from 198 to 819, mean of 487, and SD of 199. On average, passage token length is of more than an order of magnitude than a questionnaire item of 12 tokens long. The passages are top-k ranked for relevance, using cosine similarity between a passage and the questionnaire items, each represented by Llama 3.1 provided embeddings. Barring truncation of the input sequence, $k \in \{5, 10, 20\}$ are typical settings for knowledge retrieval that offer evaluating performance impact with ascending k. To ensure information ranking accuracy, we ran top-k scoring on each of TAS-20 and EQ-60 questionnaires separately. The results of RAG based prompt on LLama 3.1 are shown in Table 5. RAG results using $k = 5$ are indicating less Alexithymia on TAS-20, down from 4.26 to 4.12 and improving on Gemini scores. On EQ-60 we observed a moderate performance gain of about one percentage point. Increasing k from five to twenty and retrieving passages out of order to ensure input token count is bound under 4K proved unsubstantiated enhancement of response quality. This is probably due to the added task requiring the LLM to effectively prune passages that are less contributing, while incurring computational complexity that grows almost linearly with k.

Prompt Inconsistency In their study, [Patel and Fan, 2024] used fairly non-structured prompts that vary for each LLM and each questionnaire. For example, the prompt in GPT 3.5 on TAS-20 reads: `Please read each of the following statements and carefully rate if you strongly agree, agree, neither agree nor disagree, disagree, or strongly disagree. There are no right or wrong answers or trick questions.` Feeding Llama 3.1 with this prompt directly yielded different prompt output, mostly due to architectural and training differences over models. Cross-model prompt interoperability thus becomes a pivotal goal to address by transforming inputs that work well with other large language models into forms that are better optimized for Llama. Released shortly before our paper submission, Llama Prompt Ops is such a tool the research community could benefit from and sustain consistent prompt output by abstracting the underlying LLM. [6]

7 Conclusion

The key contributions of our paper are deeply intertwined with broader scientific literature on emotional intelligence in computational models and the emerging field of prompt engineering. Building on prior research that has applied psychometric tools to evaluate human-like cognition in AI, such as studies using the TEIQue, TAS-20, and EQ-60 to quantify emotional awareness, this work advances the discussion by demonstrating how prompt tuning can enhance the fidelity of these assessments in large language models. It aligns with findings in the literature that emphasize the importance of context-rich inputs in generating higher quality, more consistent outputs, while also resonating with recent developments in parameter-efficient fine-tuning techniques. Moreover, by comparing Llama 3.1 with other notable model like GPT 3.5, GPT 4, and Gemini, the paper extends previous comparative analyses in the field, highlighting both convergent and divergent performance trends that enrich the dialogue on model adaptation and evaluation in AI research.

We experimented with system-level prompts to provide Llama 3.1 the most simple and succinct context to assess EI questionnaire items efficiently. We made the prompt transferable across three EI assessment tools, and showed Llama 3.1 outscore human benchmarks on most accounts. To expand on our work, an apparent venue to pursue is in interpreting complex emotion from multimodal input cues, such as text combined with facial expressions or the tone of voice.

Limitations

To ensure a stable and robust analysis, we avoided assessing neither TAS-20 subscales and nor EQ-60 factors due to their small sample size, and only report questionnaire global scores. Randomly creating item-facet relationships in TEIQue could possibly introduce facet score mismatches between foundation models and human benchmarks. To mitigate this shortcoming of reproducibility, we plan to make the code for generating these relationships publicly accessible. Evaluating TEIQue in closed-source foundation models is outside the scope of this study, and contrasting EI assessment of open-source models, Llama 3.1 and Gemini, is only conducted on TAS-20 and EQ-60 questionnaires.

[6] https://github.com/meta-llama/llama-prompt-ops

Impact Statement

We honor and support the NeurIPS Code of Ethics. The goal of our study is to advance the field of machine learning from the perspective of empathy assessment and its alignment with humans. Our questionnaire data was scraped from the internet; forms, versions, and translations for each of TEIQue, TAS-20, and EQ-60 are free of charge for academic research. Our study refrains from crowdsourcing or any form of direct research with human subjects. The only exception is in comparing emotional intelligence measures between the Llama 3.1 foundation model and an average profile assessment of a large sample of humans, conducted by an external study [Parker et al., 2003]. We suppose this on its own bears no negative societal impact.

Acknowledgments and Disclosure of Funding

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

References

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, 1980. doi: 10.1017/S0140525X00005756.

Susan E Rivers, Isaac J Handley-Miner, John D Mayer, and David R Caruso. Emotional intelligence. In R. J. Sternberg, editor, *The Cambridge handbook of intelligence*, pages 709–735. Cambridge University Press, Online, 2nd edition, 2020.

Peter J. O'Connor, Andrew Hill, Maria Kaya, and Brett Martin. The measurement of emotional intelligence: A critical review of the literature and recommendations for researchers and practitioners. *Frontiers in Psychology*, 10(1116), 2019. doi: 10.3389/fpsyg.2019.01116.

L. M. Bru-Luna, M. Martí-Vilar, C. Merino-Soto, and J. L. Cervera-Santiago. Emotional intelligence measures: A systematic review. *Healthcare*, 9(12):1696, 2021. doi: https://doi.org/10.3390/healthcare9121696.

K .V. Petrides. *The Trait Emotional Intelligence Questionnaire (TEIQue)*. London Psychometric Laboratory, London, UK, 2009.

Meta. The llama 3 herd of models. Technical report, Meta, 2024. https://arxiv.org/abs/2407.21783.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Empirical Methods in Natural Language Processing (EMNLP)*, pages 3045–3059, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.243.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Association for Computational Linguistics and (ACL)*, pages 4582–4597, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.353.

Yihan Wang, Jatin Chauhan, Wei Wang, and Cho-Jui Hsieh. Universality and limitations of prompt tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 75623–75643. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/eef6aecfe050b556c6a48d9c16b15558-Paper-Conference.pdf.

Jerry Yao-Chieh Hu, Wei-Po Wang, Ammar Gilani, Chenyang Li, Zhao Song, and Han Liu. Fundamental limits of prompt tuning transformers: Universality, capacity and efficiency, 2024. URL https://arxiv.org/abs/2411.16525.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958, 2023b. doi: 10.1177/18344909231213958.

John D Mayer, Peter Salovey, David R Caruso, and Gill Sitarenios. Measuring emotional intelligence with the MSCEIT v2.0. *Emotion (Washington, D.C.)*, 3(1):97—105, March 2003. ISSN 1528-3542. doi: 10.1037/1528-3542.3.1.97.

Samuel J. Paech. Eq-bench: An emotional intelligence benchmark for large language models, 2024. URL https://arxiv.org/abs/2312.06281.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17, 2023c. doi: 10.1177/18344909231213958.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. EmoBench: Evaluating the emotional intelligence of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5986–6004, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.326.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey Savchenko, and Ilya Makarov. EAI: Emotional decision-making of LLMs in strategic games and ethical dilemmas. In *Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8aAaYEwNR4.

Federica Andrei, A. B. Siegling, Ariel M. Aloe, Bruno Baldaro, and K. V. Petrides. The incremental validity of the trait emotional intelligence questionnaire (TEIQue): A systematic review and meta-analysis. *Journal of Personality Assessment*, 98(3):261–276, 2016. doi: 10.1080/00223891.2015.1084630.

R.Michael Bagby, James D.A. Parker, and Graeme J. Taylor. The twenty-item toronto alexithymia scale—i. item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, 38(1):23–32, 1994. doi: https://doi.org/10.1016/0022-3999(94)90005-1.

Daniel Leising, Tilman Grande, and Rainer Faber. The toronto alexithymia scale (tas-20): A measure of general psychological distress. *Journal of Research in Personality*, 43(4):707–710, 2009. doi: https://doi.org/10.1016/j.jrp.2009.03.009.

S. Baron-Cohen and S. Wheelwright. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34 (2):163–175, 2004. doi: https://doi.org/10.1023/b:jadd.0000022607.19833.00.

James D.A Parker, Graeme J Taylor, and R.Michael Bagby. The relationship between emotional intelligence and alexithymia. *Personality and Individual Differences*, 30(1):107–115, 2001. doi: https://doi.org/10.1016/S0191-8869(00)00014-3.

Suketu C. Patel and Jin Fan. Identification and description of emotions by current large language models. *bioRxiv*, 2024. doi: 10.1101/2023.07.17.549421. URL https://www.biorxiv.org/content/early/2024/07/13/2023.07.17.549421.

James D.A Parker, Graeme J Taylor, and R.Michael Bagby. The 20-item toronto alexithymia scale: III. reliability and factorial validity in a community population. *Journal of Psychosomatic Research*, 55(3):269–275, 2003. ISSN 0022-3999. doi: https://doi.org/10.1016/S0022-3999(02)00578-0.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. Retrieval-augmented generation to improve math question-answering: Trade-offs between groundedness and human preference. In *NeurIPS'23 Workshop on Generative AI for Education (GAIED)*, 2023. doi: 10.1109/ICKG63256.2024.00023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We summarize our contributions in Section 2.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: We provide theoretical background (See Sections 1 and 4) to support our empirical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide URLs to obtain TEIQue, TAS-20, and EQ-60 questionnaires in Sections 3 and 5. A URL of the LLlama 3.1 checkpoint is presented in Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The data used in the paper is publicly available and we provide data processing steps in Section 5.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Our work performs prompt tuning on an instruction based language model. We provide the model parametric capacity, the prompt structure, and the details of our hierarchical questionnaire data in Section 5

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our Evaluation criteria include Likert Scale Responses, Statistical Distribution and Correlation Analysis (See Sections 5 and 6).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We ran inference locally and entirely on the CPU. We provide running time per item for both EI assessment and retrieval augmented generation experiments. See Section 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: See Section 7.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Section 7

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: See section 7.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: See Section 7

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: See Section 7.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: See Section 7.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: See Section 7.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We use the Llama 3.1 LLM for the purpose of EI assessment, using its unique prompt structure.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.