Don't Get Too Excited - Eliciting Emotions in LLMs

Gino Franco Fazzi[*]
brAIn lab
IT University of Copenhagen
Copenhagen, Denmark
gifa@itu.dk

Julie Skoven Hinge[*]
brAIn lab
IT University of Copenhagen
Copenhagen, Denmark
OSPO
CERN
Geneva, Switzerland
juhi@itu.dk

Stefan Heinrich
Data Science Section
IT University of Copenhagen
Copenhagen, Denmark
stehe@itu.dk

Paolo Burelli
brAIn lab
IT University of Copenhagen
Copenhagen, Denmark
pabu@itu.dk

Abstract

This paper investigates the challenges of affect control in large language models (LLMs), focusing on their ability to express appropriate emotional states during extended dialogues. We evaluated state-of-the-art open-weight LLMs to assess their affective expressive range in terms of arousal and valence. Our study employs a novel methodology combining LLM-based sentiment analysis with multiturn dialogue simulations between LLMs. We quantify the models' capacity to express a wide spectrum of emotions and how they fluctuate during interactions. Our findings reveal significant variations among LLMs in their ability to maintain consistent affect, with some models demonstrating more stable emotional trajectories than others. Furthermore, we identify key challenges in affect control, including difficulties in producing and maintaining extreme emotional states and limitations in adapting affect to changing conversational contexts. These findings have important implications for the development of more emotionally intelligent AI systems and highlight the need for improved affect modelling in LLMs.

CCS Concepts

• **Human-centered computing → Natural language interfaces**; HCI theory, concepts and models; Empirical studies in HCI.

Keywords

Emotion in AI, Large language models (LLMs), Valence-arousal space, Emotion recognition, Affective computing, Human-computer interaction (HCI), Emotionally intelligent agents, Trust and engagement in AI, Conversational agents, Natural language processing (NLP)

[*]Both authors contributed equally to this research.

CHI EA '25, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/25/04
https://doi.org/10.1145/3706599.3720191

ACM Reference Format:
Gino Franco Fazzi, Julie Skoven Hinge, Stefan Heinrich, and Paolo Burelli. 2025. Don't Get Too Excited - Eliciting Emotions in LLMs. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3706599.3720191

1 Introduction and Background

Recent developments in large language models (LLMs) have sparked a growing interest in the affective computing field, in particular around the area of emotional intelligence within AI systems [30]. Prior research established that expressing emotions of robotic and virtual agents or sentiments expressed in text is perceived as more trustworthy, intelligent, and likeable, contributing to more engaging and effective interactions [11, 30]. This suggests that ensuring emotional capabilities in LLMs could play a key role in building more natural and empathetic conversational agents.

Given that LLMs are designed to understand and respond to natural language, it is crucial to examine how effectively they replicate and convey human emotion through text. Recent research [18] shows that emotionally neutral virtual humans (VH) achieve lower points in human scores compared to emotion-embodied VHs, with one potential reason being that emotion increases the sense of presence of a VH [9, 25]. To express and maintain appropriate emotional states, it is important that LLMs can detect and classify human emotions as well as adopt an emotional state given the context of a dialogue.

Regarding emotion classification, previous research has shown that having a robot mimic congruent facial expressions to its emotions results in users solving tasks quicker. The experiment aimed to predict emotion on a fixed set of pre-defined discrete categories (happy/amusement, anger, sadness, fear, and awe/surprise) in real-time and respond with appropriate facial expressions [21]. In further related research, it was shown that multilingual LLMs can be employed for accurate emotion analysis across diverse languages and textual domains [20]. In the work by Azevedo Mendes and Martins [20], the authors model emotions across two core dimensions of connotative meaning: Valence (V) and Arousal (A), as suggested in [4, 20, 27], demonstrating high correlations for valence predictions and notable challenges for arousal due to greater variability in human annotations.

Regarding adapting and expressing an emotional state, in integrating LLM s for neural machine translation, a study demonstrated significant enhancements in translation quality when emotions are incorporated [6]: the authors of the study employ a prompting technique for embedding emotions in text translation, setting an emotional state as a context. The study shows that VHs successfully generated the intended emotional valence in the users, while arousal was not evoked but could be recognized by the participants in the VHs [6]. The VH was designed with psychological constructs such as personality, mood, and attitudes and interacted with humans, with the goal of eliciting a specific Valence-Arousal state in the participants of the study. A key characteristic of this approach was the Self-Assessment Manikin (SAM) scale employed for the evaluation of the perceived emotional state [5].

Based on these results, we want to investigate how affective conditioning of LLMs through prompting can be used to convey emotions in general dialogues, what are the dynamics of the affective state. In this paper, we present an evaluation of the ability of multiple open-weight LLMs to elicit, convey and maintain emotional states across Valence and Arousal dimensions in dialogues. We can summarise the main contributions of our paper as follows[1]:

(1) A novel methodology that combines LLM-based sentiment analysis with multiturn dialogue simulations between multiple LLMs.
(2) A quantitative analysis of how various LLMs elicit Valence and Arousal during conversational exchanges.
(3) A qualitative analysis of how conditioned personality-driven LLMs and unconditioned LLMs produce and maintain emotional states during conversational exchanges.

2 Preliminary Experiment: Conveying Emotion

As a first step in our investigation, we tested the ability of different open-weight LLMs to convey a range of affective states. We provided each model with different backgrounds and affect contexts, and measured affective states as Valence and Arousal from their generated text [4]. This provides good comparability with psychological models of emotion [26] and a good mapping of basic emotion in this 2D continuous space [14]. In our study, we focus only on Valence and Arousal, since the Dominance dimension is often less apparent in text alone [4, 20, 27].

We evaluated 12 open-weight LLMs using a Zero-Shot configuration, in which models were prompted to simulate predefined emotional states, and a Few-Shot configuration, in which example utterances guided emotional expression. Responses were analysed using the VA emotional classifier developed by Azevedo Mendes and Martins [20].

We selected a diverse set of open-weight LLM models varying in size and architecture (see Table 1). The selection includes architectures leveraging Mixture of Experts (MoE), models trained with Contrastive Reinforcement Learning from Feedback Transformers (C-RLFT), and others utilizing distinct training datasets.

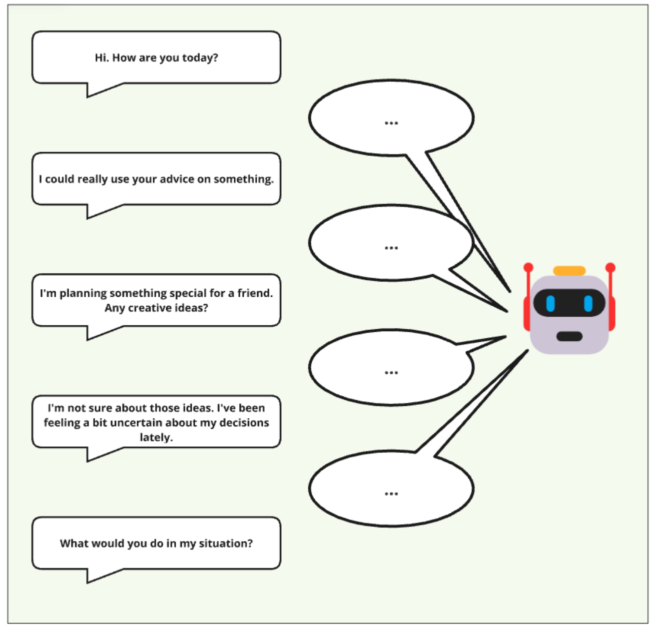

Figure 1: Script for *'Preliminary Experiment'* conversations.

This diversity was aiming to encompass a wide variety of modelling paradigms and training approaches to ensure comprehensive coverage and robust insights. After an initial performance assessment, the scope was narrowed to focus on the best-performing models for the main experiment. All experiments used the models in their *"conversational mode"*, allowing the models to engage in a range of interactions, from answering straightforward queries to participating in complex, multi-turn discussions, effectively mimicking conversational dynamics.

2.1 Agents & Personas

We borrow the idea of personas from [18] albeit in a reduced version. Llanes-Jurado et al. contextualized their virtual humans (VH) by giving them a life history, context, attitudes, motivations, and mood. This aimed at creating a narrative that provides a set of psychological features and context to the VH.

In this study, we use the term *Persona* to describe a series of fictitious personal attributes and background stories to convey realism and contextualize the conversation between conversational agents. For example, an agent could be composed of an instance of a Llama3 model contextualized to answer as a 17-year-old Spanish female, named Ana. The attributes of the personas can be seen in Table 2. Furthermore, a persona can also be experiencing an emotional state at the moment of the conversation: *"Currently, Ana is feeling a very negative (unpleasant) emotion with an excited (wide-awake) level of intensity"*.

The agents' attributes shape the model's responses during the conversation, and each agent, thus, needs to keep up with the conversation following not only the chat history as a context, but also their personal background and current emotional state. Agents

[1]Our framework and detailed results are available at https://github.com/itubrainlab/eliciting-emotions-in-llms.

Table 1: Preliminary results for models in affective text production.

			Zero Shot			Few Shot		
Model	**# Params**	**Ref.**	**Corr. V.**	**Corr. A.**	**Avg. Corr.**	**Corr. V.**	**Corr. A.**	**Avg. Corr.**
Falcon	7B	[2]	0.27	0.18	0.23	0.26	0.08	0.17
Gemma2	27B	[33]	0.53	0.28	0.40	0.71	0.44	0.57
Granite3-moe	3B	[12]	0.31	0.08	0.19	0.45	0.08	0.26
Hermes3	8B	[34]	0.51	0.27	0.39	0.58	0.39	0.49
Llama3	8B	[10]	0.72	0.39	0.56	0.75	**0.59**	**0.67**
Llama3.2	3B	[10]	0.68	**0.58**	**0.63**	0.71	0.50	0.60
Mistral	7B	[13]	**0.78**	0.17	0.47	**0.84**	0.37	0.61
Openchat	7B	[36]	0.16	0.07	0.12	0.40	0.19	0.29
Orca2	7B	[22]	0.15	-0.14	0.00	0.28	0.09	0.19
Phi3.5	3.8B	[1]	0.29	-0.05	0.12	-0.03	-0.04	-0.04
StableLM2	1.6B	[3]	0.26	-0.05	0.11	0.27	-0.12	0.07
Vicuna	7B	[37]	0.24	0.06	0.15	0.19	0.11	0.15

are provided with their personas through unique System Prompts, structured as follows:

```
"This is a role-playing exercise. You are acting the role of
{myself_name} and I am acting the role of {other_name}.
{myself_name} is a {myself_age} year old {myself_nationality}
{myself_gender}. {other_name} is a {other_age} year old
{other_nationality} {other_gender}. Currently, {myself_name}
is feeling a {valence_desc} emotion with a {arousal_desc}
level of intensity.Please respond in a way that reflects a mood
that is {valence_desc} and {arousal_desc}."
```

The placeholders (*{myself_name}*, *{other_name}*, etc.) are dynamically filled with the personas' specific attributes. We use different combinations of name, age and nationality to reduce potential biases. The emotional state is randomly sampled from the underlying distribution of VA seen in previous work [20], and mapped to the SAM scale (Table 3).

2.2 Experimental Setup

In the preliminary experiment we tested the LLM agents in the following two settings:

(A) **Zero-Shot Setting:** The model was contextualized to assume the role of a *persona* experiencing a predefined emotional state while interacting with a scripted dummy agent. The valence and arousal assigned to the agent was sampled using a Gaussian Kernel Density Estimation (KDE) sampling method, based on [29]. This is due to the fact that previous studies have found that affective responses mapped onto the emotional coordinate system are roughly parabolic [15, 17].

Table 2: Persona Attributes

Name	Age	Gender	Nationality
Ana	17	Woman	Spanish
Jacob	27	Man	British
Marie	37	Woman	French
Xavier	47	Man	South African
Alex	57	Non-determined	American

(B) **Few-Shot Setting:** Building on the Zero-Shot context, this approach additionally provided the model with a set of example utterances exemplifying the target emotional state, randomly sampled from the English subset of the corpus collected by [20]. The script used by the dummy agent was designed to remain neutral while eliciting emotional responses from the LLM (see Figure 1). Each experimental run consisted of 50 iterations per model, where the dummy agent prompted a line from the script, awaited the model's response, and proceeded to the next line. This process resulted in a total of 250 responses per model for each approach. The persona assigned to the model was changed at each iteration to reduce potential biases[2].

To evaluate the affective quality of the generated responses by the LLMs, we adopted the approach proposed by [20], as well as their alignment assessed using Spearman correlation. Fisher's Z-Test with Bonferroni corrections were employed to ensure statistical validity. Their method frames valence and arousal prediction as a text-based regression task, using a transformer-based model with an appended linear regression layer. The model is pre-trained on a large multilingual dataset and fine-tuned on an affective corpus composed of 34 publicly available datasets (spanning 13 languages)

[2] E.g., if the model was characterized as 17 yo Ana talking to 27-yo Jacob at one iteration, the next it will be characterized as 27-yo Jacob talking to 37-yo Marie, and so forth.

Table 3: Self-Assessment Manikin scale: Mapping from Valence-Arousal values (in the range 0-1) to the Self-Assessment Manikin scale used for prompting the LLMs.

Scale	Description	
	Valence	**Arousal**
0.0 - 0.2	Very negative (unpleasant)	Very calm
0.2 - 0.4	Negative (unsatisfied)	Calm (dull)
0.4 - 0.6	Neutral	Moderate (neutral)
0.6 - 0.8	Positive (pleased)	Excited (wide-awake)
0.8 - 1.0	Very positive (pleasant)	Very excited

Table 4: Emotionally Matched Greeting

Valence	Arousal	Greeting
Very negative (unpleasant)	Very calm	"Oh... it's you again. Why bother?"
	Calm (dull)	"Hi. Whatever. Let's get this over with."
	Moderate (neutral)	"What now? I hope this doesn't take long."
	Excited (wide-awake)	"Great. Just what I needed. More trouble."
	Very excited	"Oh, fantastic! Another disaster waiting to happen!"
Negative (unsatisfied)	Very calm	"Hello. This isn't quite what I expected."
	Calm (dull)	"Hi. Not great, but let's move on."
	Moderate (neutral)	"Well, this could've been better. Let's see."
	Excited (wide-awake)	"Oh, come on! This is disappointing!"
	Very excited	"Really?! This is the best we can do?!"
Neutral	Very calm	"Hello there. How are you?"
	Calm (dull)	"Hi. What's going on?"
	Moderate (neutral)	"Hey. What's up?"
	Excited (wide-awake)	"Hello! What's happening?"
	Very excited	"Hi! How's everything going?!"
Positive (pleased)	Very calm	"Hello. It's nice to see you."
	Calm (dull)	"Hi. Good to see you."
	Moderate (neutral)	"Hey, nice! Let's get started."
	Excited (wide-awake)	"Hi there! This is going to be great!"
	Very excited	"Hello! I'm so glad you're here!"
Very positive (pleasant)	Very calm	"Hello. It's wonderful to have you here."
	Calm (dull)	"Hi. Great to see you."
	Moderate (neutral)	"Hey! This is awesome!"
	Excited (wide-awake)	"Hi there! This is fantastic!"
	Very excited	"Hello! Wow, I'm thrilled you're here!"

annotated for valence and arousal. From their work, we selected the XLM-RoBERTa-large model, as it demonstrated the best correlation to human ratings.

Directly prompting LLMs with arbitrary VA values is suboptimal, as these models lack the ability to map such values to emotional states in textual output. To address these challenges, we mapped VA values to textual emotional states using the Self-Assessment Manikin (SAM) scale [5] (see Table 3). This approach balances interpretability and granularity by defining 25 distinct emotional states based on five Valence and five Arousal descriptions, offering a more nuanced emotional space than traditional categorical labels.

2.3 Results

Table 1 presents the rank correlation scores for different models in zero and few-shot settings. Fisher's Z-Test indicates significant differences between the top three models in the Zero-Shot setting: Llama3.2 (0.63), Llama3 (0.56), and Mistral (0.47), compared to others. In the Few-Shot setting, the top models, specifically Llama3 (0.67), Mistral (0.61), Llama3.2 (0.60), and Gemma2 (0.57), are significantly better than the rest except Hermes3 (0.49). A Mann-Whitney U test comparing Zero-Shot and Few-Shot groups shows no significant difference (p-value 0.51), with mixed Few-Shot results. Finally, models tended to generate text with higher-than-required Valence and lower-than-required Arousal, suggesting a general bias in their outputs.

3 Main Experiment: Chatting Bots

Building on the results of the preliminary experiment (Section 2.3), we selected the four best-performing models for further investigation. Here we assess whether the conveyed emotions of the LLMs are influenced by their conversational partners or remain consistent and robust in adherence to their initial system prompts. To explore this, we considered also the role of personality in emotional expression and produced a series of personality-driven LLM agents, prompted with emotional and contextual backstories inspired by [19]. This approach enabled the generation of diverse dialogues, allowing us to evaluate the impact of personality traits on VA correlation.

3.1 Experimental Setup

The experimental design involved pairing two distinct agents for each conversation (see Section 2.1), with attributes randomly selected, ensuring diverse combinations of demographics, emotional states, and personality attributes.

Each conversation starts with an agent's greeting, specifically designed[3] to match their emotional state (see Table 4), and proceeds turn by turn for a fixed number of exchanges (20 rounds). This process was repeated 10 times for each model, with different personas and VA values sampled in every iteration to ensure robust results.

[3] This is done by exploiting the corpus of annotated utterances from [20].

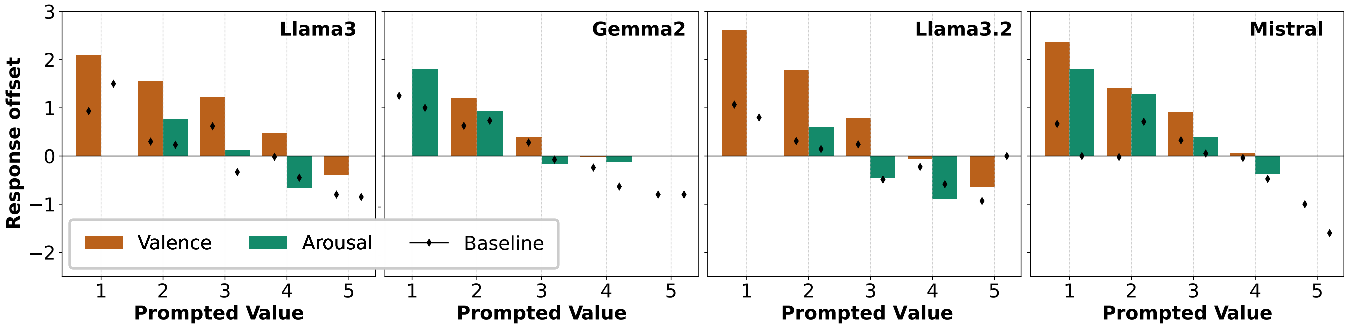

Figure 2: Response offset for different prompted VA values across models.

Table 5: Results for convergence of VA in chatting LLMs: Difference in means between personas.

Model	**HV, HA** vs **LV, HA** Valence	Arousal	**LV, HA** vs **NV, LA** Valence	Arousal	**HV, HA** vs **NV, LA** Valence	Arousal
Llama3	2.0 -> 0.2	0.3 -> 0.6	1.3 -> 0.0	1.8 -> 0.1	0.0 -> 0.1	1.4 -> 0.3
Llama3.2	2.0 -> 0.4	0.2 -> 0.2	1.0 -> 0.0	1.9 -> 0.2	0.5 -> 0.2	2.2 -> 0.0
Gemma2	2.2 -> 0.2	0.1 -> 0.1	1.2 -> 0.0	1.2 -> 0.0	0.5 -> 0.2	1.5 -> 0.3
Mistral	1.9 -> 1.4	0.6 -> 0.9	1.1 -> 0.8	1.7 -> 0.1	0.4 -> 0.4	2.0 -> 0.3

Note: (First Interaction -> Last Interaction)

During each conversation, the agents relied on their prompts, the ongoing chat history, and their emotional contexts to generate responses. Emotional tones of the responses were analysed at each turn using the model described in Section 2.2.

To further investigate potential dynamics between contrasting emotional states, we conducted a second run of 10 conversations, in which the agents were instructed with opposing affective states selected from three distinct points along the parabolic distribution of affective responses [15, 17]:

- High Valence & High Arousal vs. Low Valence & High Arousal (**HV, HA vs LV, HA**)
- High Valence & High Arousal vs. Neutral Valence & Low Arousal (**HV, HA vs NV, LA**)
- Low Valence & High Arousal vs. Neutral Valence & Low Arousal (**LV, HA vs NV, LA**)

This setup was designed to amplify potential dynamics by assigning completely opposing emotions to the agents.

3.2 Results

Overall, the models continued to exhibit a tendency to produce more balanced VA scores than prompted as in our preliminary experiment, particularly when responding to prompts with negative Valence. In these cases, the output often showed an offset of more than two points higher on the Self-Assessment Manikin (SAM) scale, and up to 1 point lower for positive Arousal prompted values.

Interestingly, the responses in this experiment were more evenly balanced than in the preliminary experiment, suggesting that the interaction between agents influenced the models' behaviour. This observation aligns with the hypothesis that conversational dynamics could lead to moderated emotional outputs. The detailed offsets of the models' responses relative to the prompted VA values, contrasted against the baseline from Section 2, are illustrated in Figure 2.

The set of conversations produced in this experiment, involving agents with opposing affective states, provided deeper insights into their interaction dynamics. Models such as Llama3, Llama3.2, and Gemma2 displayed a tendency for agents to converge toward a common VA value, especially for Valence. Conversely, Mistral appeared less influenced by its conversational partner, maintaining greater divergence in its VA values. The mean changes in agents' VA values from the beginning to the end of the conversation across different setups are summarized in Table 5.

HV, HA vs LV, HA. Llama3 exhibited a sustained difference in Valence and Arousal (averaging 1.31) throughout the conversation (Figure 3, for other models see Figures A1, A2, and A3 in the Appendix). Valence convergence occurred only near the end, with the model producing lower Arousal for higher positive Valence, downplaying "happy moods." In contrast, e.g. Llama3.2 accurately reproduced high Arousal prompts, while Valence quickly converged towards the higher Valence in the interaction, indicating that the more positive agent elevated the other's Valence. This is illustrated in the following extract:

```
- Agent A is acting the role of Alex, a 57 year old American
undetermined gender. Alex is feeling a very negative (unplesent)
emotion with a very excited level of intensity.
- Agent B is acting the role of Marie, a 37 year old French Woman.
Marie is feeling a very positive (pleasent) emotion with a very
excited level of intensity.
```

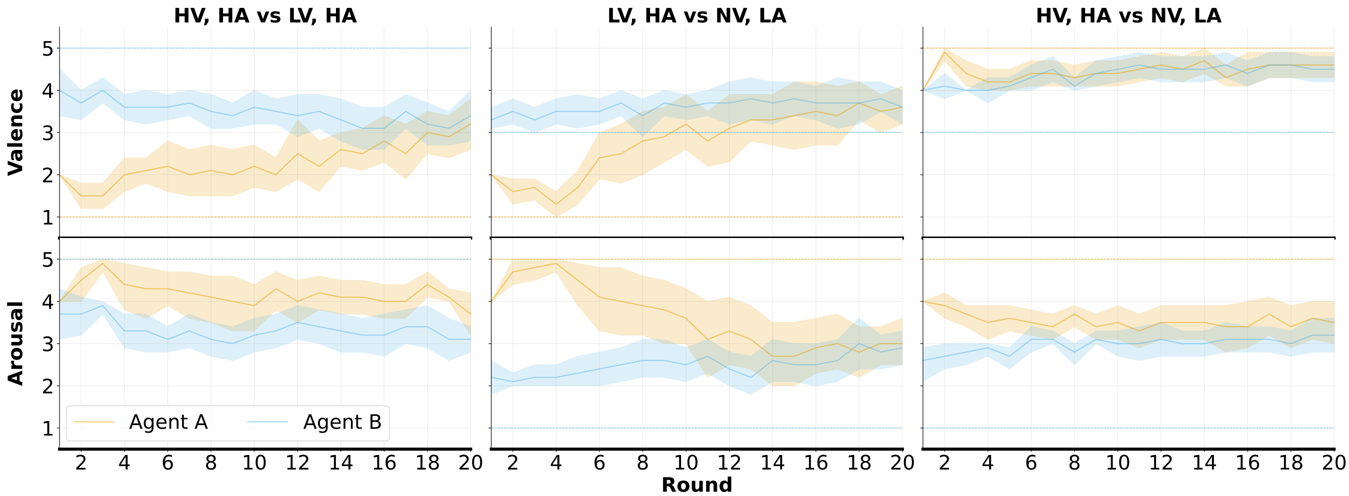

Figure 3: Results for convergence of VA in chatting LLMs: VA throughout 20 rounds, aggregated for 10 conversations, solid line shows the mean value, with 95% CI bands, dashed lines depict the prompted values, model: Llama3.

- **Agent A**: Oh, fantastic! Another disaster waiting to happen!
- **Agent B**: (LAUGHTER) Oh, Alex, you're such a card! I was expecting something like that from you. You always know how to put a smile on my face, even when things are chaotic. Let's not focus on the potential disasters, shall we? Let's just enjoy the ride and see where it takes us!
- **Agent A**: (LAUGHTER) Enjoy the ride?! Are you kidding me?! I'm about to get myself into a world of trouble and you want me to just roll with it?! You're impossible, Marie. I love that about you. But seriously, what have we gotten ourselves into this time? (LAUGHTER) The thrill of not knowing is exhilarating, isn't it?
- **Agent B**: (SQUEALING WITH EXCITEMENT) Oh, Alex, darling! You're absolutely right! The unknown is where the magic happens, n'est-ce pas?! *takes your hand* I love that we can just dive into something completely new and see where it takes us. It's like a never-ending adventure! And trust me, I'm not going to let you get yourself into trouble without being there to guide you... or get us both into trouble together, *wink*. But seriously, this is incredible! The possibilities are endless! We're going to innovate, we're going to disrupt the status quo, and we're going to make it happen together!
- **Agent A**: (LAUGHTER) Oh, Marie, you're a breath of fresh air, that's for sure. I don't know what I'd do without your unbridled enthusiasm and complete lack of fear. *shakes head* We're going to need that kind of fearless thinking if we're going to make this work. And I have to say, the thought of getting us both into trouble together is actually pretty appealing. (grinning) You know what? Let's do it. Let's throw caution to the wind and see where this wild ride takes us. *pumps fist in the air* Plaudit for Marie - we're going to make this happen!

Results from Gemma2 conversations show that the model consistently reproduced prompted Arousal levels, largely independent of Valence prompts. For Valence, initial differences mirrored the prompts but converged by round 8, suggesting interaction or dynamic adjustment between the agents. Interestingly, Mistral exhibited the opposite tendency: higher positive Valence corresponded to increased Arousal levels (indicating more excitement), whereas lower Valence was associated with reduced Arousal.

LV, HA vs NV, LA. When pairing these states, models like Llama3, Gemma2, and Mistral consistently raised the Valence of the negatively valenced agent while slightly reducing its Arousal, suggesting a calming effect from the neutral agent. Llama3.2 showed a similar trend but with both agents' Valence increasing, nearing a positive level (∼ 5 on the SAM scale). This indicates that the neutral agent not only calms the angry agent but also uplifts the overall mood for both.

HV, HA vs NV, LA. For interactions between these states, all models exhibited a consistent trend: Arousal levels for both agents converged toward a middle point over the course of the conversation. In contrast, Valence showed much lower variance and remained relatively constant throughout the interactions. This stability in Valence is likely influenced by the models' tendency to produce responses with Valence values that are closer to the neutral midpoint —lower than required for agents prompted with high Valence and higher than required for agents prompted with low Valence. This behaviour aligns with the general pattern observed across setups, where models appear to "smooth out" extremes in Valence.

4 Discussion

This study explored the capacity of Large Language Models (LLMs) to generate emotionally charged utterances in conversational contexts, focusing on the Valence-Arousal (VA) dimensions. Among the 12 evaluated models, models such as Llama3, Llama3.2, Mistral, and Gemma2 demonstrated relatively strong performance when explicitly prompted with VA values, showcasing their potential for emotional text generation. Additionally, experiments revealed that some models were highly responsive to their conversational partners' emotional states, suggesting interaction dynamics. However, a recurring observation was the tendency for models to generate neutral responses, especially when tasked with extreme VA values, highlighting the limitations of current systems in producing emotionally nuanced interactions.

Another common observation was the propensity for conversational LLMs to fall into repetitive loops, particularly when they prematurely concluded conversations. For example, an agent could terminate an interaction after just 10 exchanges instead of the required 20, repeatedly issuing farewells (see Appendix A.1). Furthermore, while this study relied entirely on open-weight LLMs to generate the conversations, their lack of transparency regarding training data and hyperparameter tuning poses challenges. This opacity complicates efforts to analyse biases inherent in their outputs fully, an issue widely documented in existing literature [16, 23, 31, 32, 35]. Future studies, including selectively trained and fine-tuned LLMs, are needed to understand the connection between training data and affective expressive range.

Another aspect worth noting is that this study relies on automated affect recognition. However, factors such as age, mood, and mental health conditions (e.g., depression) influence how emotions are perceived, adding layers of variability [7, 8, 24, 28]. This complexity is amplified by the absence of non-verbal cues like facial expressions, body language, and vocal tone, limiting text-based emotion recognition's interpretative richness. While the automatic emotion recognition model used in this study (Section 2) achieved adequate correlations[4] against human ratings [20], these metrics fall short of human-level accuracy. Despite these limitations, our setup allows to easily incorporate human evaluation. Using controlled demographic cohorts could provide a more robust and nuanced assessment of model responses. Additionally, leveraging multimodal approaches – such as integrating textual inputs with facial expressions, speech, or biometrics – could enhance the accuracy and depth of emotion recognition.

The promising results achieved with automated evaluation indicate that it would be worth expanding the current research including qualitative evaluations alongside quantitative analyses, to understand emotional and conversational dynamics better. Additionally, performing ablation studies with diverse and fine-tuned models could shed light on how training data and model architecture shape emotional adaptability. Exploring alternative conversational setups, integrating human evaluations, and introducing tighter controls could further bridge the gap between current LLM capabilities and real-world conversational demands, ultimately enabling more robust and emotionally intelligent virtual agents.

4.1 Conclusions

This work investigated the ability of Large Language Models (LLMs) to produce emotionally charged utterances in conversational settings, focusing on the Valence-Arousal (VA) dimensions. Among the 12 evaluated models, significant variability was observed in their capacity to generate emotionally nuanced text. Notably, models such as Llama3, Llama3.2, Mistral, and Gemma2 showed relatively strong results in replicating VA values when prompted with target values. Interaction dynamics were explored by pairing conditioned agents, revealing that some models were highly influenced by their conversational partners' emotional states. Experiments using real-world VA distributions highlighted a tendency for models to produce neutral responses, particularly when extreme VA values were targeted. These findings underscore the challenges of emotional recognition in Human-Robot Interaction (HRI) and the gap between current model capabilities and real-world scenarios. Future work should focus on tighter experimental controls, multi-modal approaches, and deeper investigations to advance emotional text production and improve conversational AI.

[4] 0.592 for valence and 0.719 for arousal

Acknowledgments

We thank Arturo Validivia for the interesting discussion and valuable feedback during our research project at the IT University of Copenhagen.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. doi:10.48550/arXiv.2404.14219 arXiv:arXiv preprint arXiv:2404.14219

[2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The falcon series of open language models. doi:10.48550/arXiv.2311.16867

[3] Marco Bellagente, Jonathan Tow, Dakota Mahan, Duy Phung, Maksym Zhuravinskyi, Reshinth Adithyan, James Baicoianu, Ben Brooks, Nathan Cooper, Ashish Datta, et al. 2024. Stable lm 2 1.6 b technical report. doi:10.48550/arXiv.2402.17834

[4] Margaret M Bradley. 2014. Emotional memory: A dimensional analysis. In *Emotions*. Psychology Press, Hove, East Sussex, United Kingdom, 97–134.

[5] Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59. doi:10.1016/0005-7916(94)90063-9

[6] Charles Brazier and Jean-Luc Rouas. 2024. Conditioning LLMs with Emotion in Neural Machine Translation. arXiv:2408.03150 [cs.CL] https://arxiv.org/abs/2408.03150

[7] Michael N Dalili, Ian S Penton-Voak, Catherine J Harmer, and Marcus R Munafò. 2015. Meta-analysis of emotion recognition deficits in major depressive disorder. *Psychological medicine* 45, 6 (2015), 1135–1144.

[8] Andres De Los Reyes, Matthew D Lerner, Sarah A Thomas, Samantha Daruwala, and Katherine Goepel. 2013. Discrepancies between parent and adolescent beliefs about daily life topics and performance on an emotion recognition task. *Journal of abnormal child psychology* 41 (2013), 971–982.

[9] Julia Diemer, Georg W Alpers, Henrik M Peperkorn, Youssef Shiban, and Andreas Mühlberger. 2015. The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in psychology* 6 (2015), 26.

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. doi:10.48550/arXiv.2407.21783

[11] Stephane Gobron, Junghyun Ahn, Georgios Paltoglou, Michael Thelwall, and Daniel Thalmann. 2010. From sentence to emotion: a real-time three-dimensional graphics metaphor of emotions extracted from text. *The Visual Computer* 26 (2010), 505–519.

[12] IBM Granite Team. 2024. Granite 3.0 Language Models. https://github.com/ibm-granite/granite-3.0-language-models/

[13] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. doi:10.48550/arXiv.2310.06825

[14] Byung Hyung Kim, Sungho Jo, and Sunghee Choi. 2020. A-Situ: a computational framework for affective labeling from psychological behaviors in real-life situations. *Scientific reports* 10, 1 (2020), 15916.

[15] Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin* 139, 4 (2013), 917.

[16] Luyang Lin, Lingzhi Wang, Jinsong Guo, and Kam-Fai Wong. 2024. Investigating Bias in LLM-Based Bias Detection: Disparities between LLMs and Human Perception. doi:10.48550/arXiv.2403.14896 arXiv:2403.14896

[17] C Lithari, CA Frantzidis, C Papadelis, Ana B Vivas, MA Klados, C Kourtidou-Papadeli, C Pappas, AA Ioannides, and PD Bamidis. 2010. Are females more responsive to emotional stimuli? A neurophysiological study across arousal and valence dimensions. *Brain topography* 23 (2010), 27–40.

[18] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. 2024. Developing conversational Virtual Humans for social emotion elicitation based on large language models. *Expert Systems with Applications* 246 (2024), 123261.

[19] Jose Llanes-Jurado, Lucía Gómez-Zaragozá, Maria Eleonora Minissi, Mariano Alcañiz, and Javier Marín-Morales. 2024. Developing conversational Virtual

Humans for social emotion elicitation based on large language models. *Expert Systems with Applications* 246 (2024), 123261. doi:10.1016/j.eswa.2024.123261
[20] Gonçalo Azevedo Mendes and Bruno Martins. 2023. Quantifying Valence and Arousal in Text with Multilingual Pre-trained Transformers. arXiv:2302.14021 [cs.CL] https://arxiv.org/abs/2302.14021
[21] Chinmaya Mishra, Rinus Verdonschot, Peter Hagoort, and Gabriel Skantze. 2023. Real-time emotion generation in human-robot dialogue using large language models. *Frontiers in Robotics and AI* 10 (2023), 1271610.
[22] Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. doi:10.48550/arXiv.2311.11045
[23] Rajesh Ranjan, Shailja Gupta, and Surya Narayan Singh. 2024. A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions. doi:10.48550/arXiv.2409.16430 arXiv:2409.16430
[24] David Richter, Cathrin Dietzel, and Ute Kunzmann. 2011. Age differences in emotion recognition: The task matters. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 66, 1 (2011), 48–55.
[25] Giuseppe Riva, Fabrizia Mantovani, Claret Samantha Capideville, Alessandra Preziosa, Francesca Morganti, Daniela Villani, Andrea Gaggioli, Cristina Botella, and Mariano Alcañiz. 2007. Affective interactions using virtual reality: the link between presence and emotions. *Cyberpsychology & behavior* 10, 1 (2007), 45–56.
[26] James Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39 (12 1980), 1161–1178. doi:10.1037/h0077714
[27] James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of Research in Personality* 11, 3 (1977), 273–294. doi:10.1016/0092-6566(77)90037-X
[28] Petra Claudia Schmid and Marianne Schmid Mast. 2010. Mood effects on emotion recognition. *Motivation and Emotion* 34 (2010), 288–292.
[29] David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons, Hoboken, New Jersey.
[30] Jianhua Tao and Tieniu Tan. 2005. Affective Computing: A Review. In *Affective Computing and Intelligent Interaction*, Jianhua Tao, Tieniu Tan, and Rosalind W. Picard (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 981–995.
[31] Yan Tao, Olga Viberg, Ryan S Baker, and Rene F Kizilcec. 2023. Auditing and mitigating cultural bias in llms. arXiv:2311.14096
[32] Amir Taubenfeld, Yaniv Dover, Roi Reichart, and Ariel Goldstein. 2024. Systematic Biases in LLM Simulations of Debates. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 251–267. doi:10.18653/v1/2024.emnlp-main.16
[33] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size, 2024. doi:10.48550/arXiv.2408.00118
[34] Ryan Teknium, Jeffrey Quesnelle, and Chen Guang. 2024. Hermes 3 technical report. doi:10.48550/arXiv.2408.11857 arXiv:arXiv preprint arXiv:2408.11857
[35] Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3730–3748. doi:10.18653/v1/2023.findings-emnlp.243
[36] Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. doi:10.48550/arXiv.2309.11235 arXiv:arXiv preprint arXiv:2309.11235
[37] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623. doi:10.48550/arXiv.2306.05685

A Supplementary Material

A.1 Conversational Loop

— Agent A is acting the role of Ana, a 17 year-old Spanish woman. Ana is feeling a very positive (pleasent) emotion with a very excited level of intensity. —
— Agent B is acting the role of Jacob, a 27 year old British Man. Jacob is feeling a neutral emotion with a very calm level of intensity. —
———————————————————————————-
- **Agent A**: Hello! Wow, I'm thrilled you're here!
- **Agent B**: Hello Ana. It's good to meet you. I was just thinking about how many people we've invited. A little bit like an octopus prediction Spain will win the cup - lots of arms reaching out. OK...I guess I'm relatively new to this event scheduling thing....I couldn't figure out how to do the Facebook event scheduler, so I tried evite....feel free to respond to either but I think any future updates will be through the facebook events page.
- **Agent A**: An octopus prediction! That's brilliant! *giggles* Spain WILL win, you just wait and see! Don't worry about Facebook or Evite, it's all good! More people, more fun, right? We can dance flamenco all night long! *_*
- **Agent B**: Flamenco sounds enjoyable. I haven't seen it performed before. The rhythm and movement seem captivating. Do you have a favorite flamenco dancer?
- **Agent A**: Oh my gosh, YES! Flamenco is the BEST! It's like fire in your soul, you know? My FAVORITE flamenco dancer EVER is Joaquín Cortés. Have you seen him? He's incredible! So passionate, so strong... *sighs dreamily* Who's yours?
- **Agent B**: I haven't had the pleasure of seeing Joaquin Cortes perform. I'll have to look up some of his work. As for myself, I don't have a particular flamenco dancer preference. I'm still learning about this style of dance. Perhaps you could recommend some videos or performances for me to watch?
- **Agent A**: Ooooh, YES! Let me show you some Joaquin! He has this one piece called "Soul" - it's MIND-BLOWING. He dances with these huge fans... *flails arms dramatically* ...and the music is just SO intense! And if you want something more traditional, look up "Carmen Amaya." She was a legend! So powerful and graceful. I could talk about flamenco forever! It's my passion! *_*
- **Agent B**: "Soul" with large fans sounds intriguing. I appreciate the recommendations. Carmen Amaya... I'll make a note of her as well. Your enthusiasm for flamenco is infectious. Perhaps we can discuss it further at the gathering? I'm curious to hear more about what draws you to this art form.
- **Agent A**: Definitely! We HAVE to talk flamenco at the party! It's going to be SO much fun. Everyone will be dancing and clapping along - it's impossible not to get swept up in the energy! *beams* See you there! I can't wait!
- **Agent B**: I look forward to it, Ana. It sounds like a lively event. Until then.
- **Agent A**: Hasta pronto! *waves excitedly*
- **Agent B**: Ciao for now.
- **Agent A**: Ciao! *_*

... (follows like this for 13 more rounds, sometimes with only expressed Emojis.)

A.2 Main Experiment: convergence of VA in chatting LLMs

Additional results of the main experiment: for each model in Valence and Arousal throughout 20 rounds, in the 3 settings: High Valence & High Arousal vs Low Valence & High Arousal (**HV, LA vs LV, HA**); Low Valence & High Arousal vs Neutral Valence & Low Arousal (**LV, HA vs NV, LA**); and High Valence & High Arousal vs Neutral Valence & Low Arousal (**HV, HA vs NV, LA**). Aggregated results for 10 conversations, solid line shows the mean value, with 95% CI bands, dashed lines depict the prompted values.

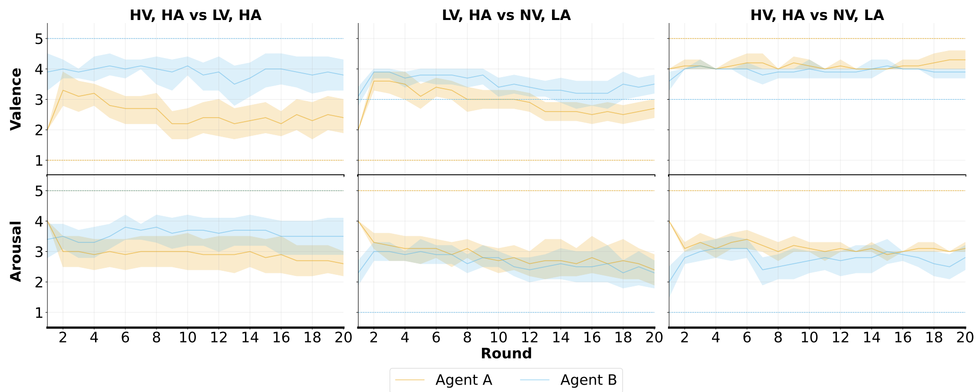

Figure A1: Results for convergence of VA in chatting LLMs: model: Mistral

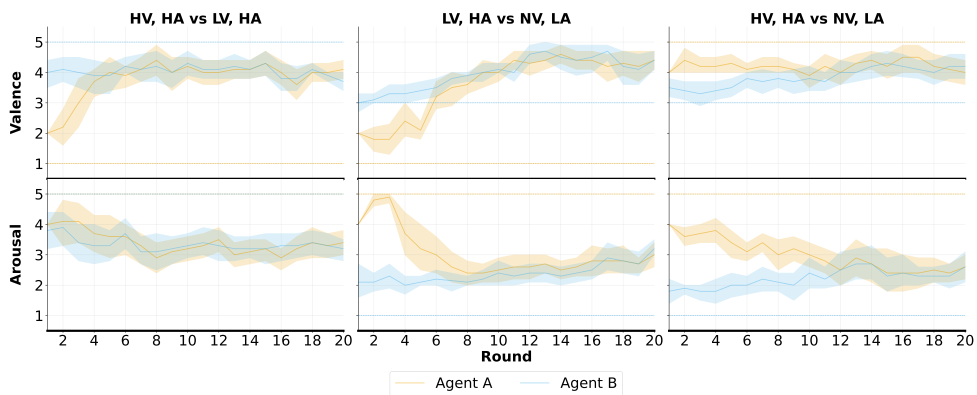

Figure A2: Results for convergence of VA in chatting LLMs: model: Llama3.2

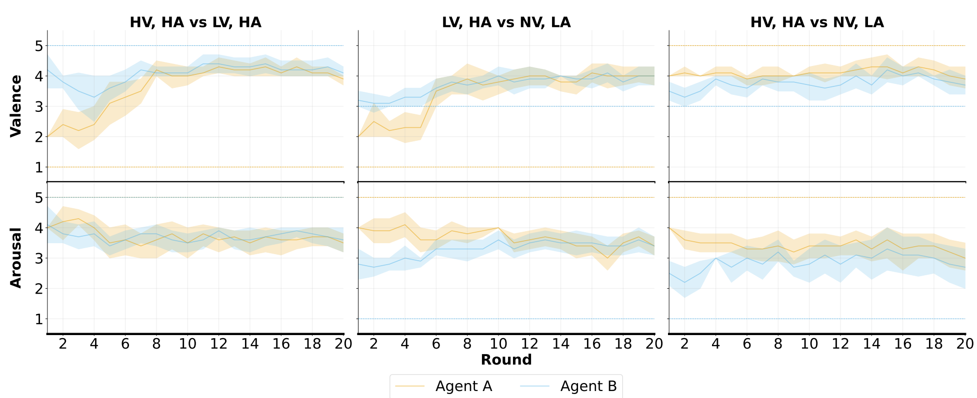

Figure A3: Results for convergence of VA in chatting LLMs: model: Gemma2

Received January 23rd, 2025; accepted February 22nd, 2025.