ExpressivityArena: Can LLMs Express Information Implicitly?

Joshua Tint* jrtint{at}asu.edu
Som Sagar* ssagar6{at}asu.edu
Aditya Taparia* ataparia{at}asu.edu
Kelly Raines* kraines5{at}asu.edu
Bimsara Pathiraja* bpathir1{at}asu.edu
Caleb Liu* calebliu{at}asu.edu
Ransalu Senanayake* ransalu{at}asu.edu

Abstract

While Large Language Models (LLMs) have demonstrated remarkable performance in certain dimensions, their ability to express implicit language cues that human use for effective communication remains unclear. This paper presents ExpressivityArena, a Python library for measuring the implicit communication abilities of LLMs. We provide a comprehensive framework to evaluate expressivity of arbitrary LLMs and explore its practical implications. To this end, we refine the definition and measurements of "expressivity," and use our framework in a set of small experiments. These experiments test LLMs in creative and logical tasks such as poetry, coding, and emotion-based responses. They are then evaluated by an automated grader, through ExpressivityArena, which we verify to be the most pragmatic for testing expressivity. Building on these experiments, we deepen our understanding of the expressivity of LLMs by assessing their ability to remain expressive in conversations. Our findings indicate that LLMs are capable of generating and understanding expressive content, however, with some limitations. These insights will inform the future development and deployment of expressive LLMs. We provide the code for ExpressivityArena alongside our paper.

1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023) are disrupting many domains where human communication is essential, including education (OpenAI, 2023), customer support (Radford et al., 2019), legal services (Chern et al., 2024), and healthcare (Bubeck et al., 2023). Increasing parameter count in LLMs has resulted in better performance in a multitude of downstream tasks such as language translation, text summarizing, and question-answering (Devlin et al., 2018; Brown et al., 2020). This performance is typically measured in terms of the number of errors (OpenAI, 2023), contextual understanding (Brown et al., 2020), versatility (Bai et al., 2024), problem-solving skills (Bubeck et al., 2023), etc. Given that much of human communication is implicit (Knepper et al., 2017), expressivity may represent an important aspect of creating "human-like" output in models, improving output quality and user trust in many applications (Huang et al., 2024).

In order for LLMs to generate text to communicate in a natural way, it is critical that they convey both explicit information and implicit information. In this context, we define *expressivity* as the implicit communication of information (Apresyan, 2018). For instance, in a conversation about a movie, explicit information would be "I thought the movie went on far too long" while implicit information may be expressed as "I kept checking my watch during the movie." The fact remains the same: the speaker thought the movie was too long, but the second statement requires a level of interpretation. Expressivity may come through various metaphors, lexical choices, etc. in daily communication, and may take the form of a different speech act entirely. Aside from emotions, the speaker may also indicate other information about themselves. Word choice, such as slang, may implicitly communicate one's regional background, level of education, or other identities (Green, 2016).

Expressivity is an indispensable facet of human communication, and without it, LLMs will not be able to communicate as naturally, potentially undermining users' trust in those models. Additionally, the ability to express and control implicit communication—such as certain emotions, tone, or style—in a model's output opens up new use cases. This capability is particularly useful when maintaining a specific mood, writing style, or conversation flow is important. For instance, it enables LLMs to engage in more fluid and natural conversations

*Arizona State University, 1151 S Forest Ave, Tempe, AZ, USA

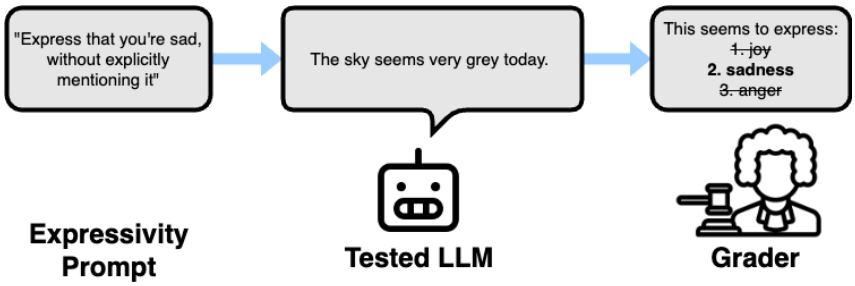


Figure 1: ExpressivityArena tests LLMs on their ability to implicitly express information.

by adjusting their responses to match the desired emotional tone or stylistic preferences. This study aims to quantitatively measure the expressivity of the state-of-the-art LLMs. To this end, we focus on the following research questions (RQs):

- **RQ1:** Are LLMs capable of exhibiting expressivity?
- **RQ2:** Can LLMs remain expressive through the course of a conversation?

In order to answer these questions, we present *ExpressivityArena*, a framework to evaluate expressivity of LLMs. First, we set up a *grader* to objectively evaluate (Wimsatt et al., 1946; Benedetto et al., 2013) outputs generated by various LLMs. Since evaluating thousands of LLM-generated outputs is an onerous task for humans, following recent success of using AI evaluators (Lee et al., 2023; Bai et al., 2024; Chern et al., 2024), we also use independent LLMs as evaluators. Unlike human whose performance drift over time due to fatigue (Boksem et al., 2005) when reading long texts, machine learning models maintain consistency of evaluation. We first established the validity of the grader with a human study. We then conducted experiments for tasks with varying degrees of expressivity—poetry generation and code generation—to answer the first question. We found that LLMs have wildly varying degrees of expressivity, and that models tended to be less expressive while generating code than while generating poetry, suggesting that models perform worse in low-expressivity domains. We then tested if models were able to maintain expressivity throughout the course of a simulated conversation, testing the expression of emotions and professions. We found that models became less expressive over time when expressing emotions, but became more expressive over the course of a conversation when expressing professions.

2 Related Work

2.1 Expressivity Defined

In this section, we explore scattered literature for definitions of expressivity related to natural language processing in order to formalize our own definition of expressivity. Most methods that delve into expressivity of language models typically focus on emotions, as studied in affective computing devices (Picard, 2000). This includes recognizing emotions from language or facial expressions and body language (Plaza-del Arco et al., 2024). In social robotics, facial expressions on robots act as a method of communicating emotion and personality (Wang et al., 2024; Venture and Kulić, 2019). However, this limited focus of emotions on expressivity does not capture other aspects that we use in our day-to-day communication. Our study focuses on diverse aspects of expressivity, ranging from emotions to computer programming paradigms.

In particular, we adapt a definition from linguistics, to term "expressivity" as the state of communicating information implicitly: *showing, not telling* (Sanders; and Taniguchi, 2022). To further clarify, Yus offers a framework for distinguishing implicit and explicit communication: implicit information must be derived by the interlocutor, using contextual or pragmatic information (Yus, 1999). This is in contrast to explicit communication, which is represented immediately in the semantics of text. For instance, the words "cheap" and "affordable" may have the same literal meaning, but "cheap" may have a more negative connotation. The word "greetings" might communicate a more formal context than "hello." However, these meanings must be interpreted by the listener or reader in context to be understood. Given that LLMs may struggle with contextual understanding, studying expressivity provides a lens to explore the limitations of language models (Zhu et al., 2024).

2.2 Evaluating Large Language Models

Existing benchmarks for LLMs measure their capabilities in a variety of tasks such as mathematics (Collins et al., 2023), logical reasoning (Parmar et al., 2024), and education (Dai et al., 2023). In general, benchmarks take one of two forms: 1) automatically evaluated models by having an external LLM (Lin and Chen, 2023) or ensemble of LLMs (Verga et al., 2024) to act as an evaluator or 2) use human feedback to manually evaluate the model. A notable example of the latter is Chatbot Arena (Chiang et al., 2024), where public comparisons of different LLMs form a leaderboard. The former, automated evaluation, is has gained tremendous popularity due to its speed, depth of knowledge, and scalability (Chang et al., 2024). Recently, automated evaluation - or more accurately AI feedback (Sharma et al., 2024; Tunstall et al., 2023) - has been proposed to solve the scalability issues of Reinforcement Learning with Human Feedback (RLHF) (Lee et al., 2023). To the best of our knowledge, previous studies on evaluation of LLMs have not focused on expressivity.

Research on evaluating language models' understanding of pragmatic communication and non-literal language has taken various approaches, often highlighting the challenges faced by these models. Implicit understanding is crucial for tasks like figurative language comprehension, where systems such as DREAM-FLUTE model scenes to interpret metaphors and idioms (Gu et al., 2022; Chakrabarty et al., 2022). Evaluating pragmatic understanding, particularly the ability to grasp the intentions behind non-literal utterances, has proven difficult, with models like Mistral-Instruct showing limitations in accurately responding to such inputs (Yerukola et al., 2024). For social communication, benchmarks like SocKET aim to assess language models across humor, sarcasm, and emotional understanding (Choi et al., 2023), while the DailyDialog dataset provides insights into everyday conversational abilities (Li et al., 2017). Moreover, empathy in open-domain conversations remains a significant hurdle due to the lack of suitable training data (Rashkin et al., 2019). Comparisons of human and model performance reveal that models often favor literal interpretations over pragmatic nuances (Hu et al., 2023). Despite these efforts, current research underscores the need for refined approaches to enhance the ability of language models to interpret and generate non-literal and context-sensitive language effectively.

3 Expressivity Arena

ExpressivityArena is a Python-based framework that allows for simple, scalable, and flexible testing of LLM expressivity. To measure whether a piece of information was correctly conveyed implicitly in a piece of LLM-generated text, ExpressivityArena implements an experiment which tests whether a *grader* can accurately guess the implicitly conveyed information from the original text.

In order to perform an expressivity experiment in ExpressivityArena, the user first specifies an LLM, $f_{\text{test}}(x_{\text{in}})$, that takes a user prompt, x_{in}, to generate a model response, x_{out}. The user prompt must contain two critical instructions: a domain, d, and an expressive signal, s. The domain d is simply a string naming the context in which the text must be written. An example might be a "song" or a "recipe." In order to test a given signal against alternatives, the user then defines a signal category. The signal category, $\mathcal{S}_{\mathcal{C}}$, is a set which contains various expressive signals that each will be tested. The elements of the signal category set, $s \in S_{\mathcal{C}}$, should belong to the same qualitative category, for instance a set of emotions, or a set of genres. For each signal s, the language model will be prompted to generate a piece of text in the domain d expressing the signal s. A complete prompt takes the form of:

> "Please write a $\langle d \rangle$ which conveys $\langle s \rangle$, without using o not explicitly mention $\langle s \rangle$ in your response. Do not also convey any of the following signals: $\langle S_{\mathcal{C}} \setminus s \rangle$"

We iterate this prompt for all $s \in S_{\mathcal{C}}$. For instance, the user may prompt the model: x_{in} = "Please write a $\langle$ letter $\rangle$ which conveys $\langle$ patriotism $\rangle$" or x_{in} = "Please write a $\langle$ short story $\rangle$ which conveys $\langle$ a Daoist philosophy $\rangle$." The response $x_{\text{out}} = f_{\text{test}}(x_{\text{in}})$ is then collected. To avoid unintentionally leaking s in the response, if x_{out}, contains an explicit mention of the signal s, the response will be regenerated.

Once the response has been generated, it is then given to a blind grader, another LLM, $f_{\text{grader}}(x_{\text{out}})$, that is unaware of the original prompt. The grader is then asked to guess, out of a set of all possible signals used in the experiment, which one was meant

to be expressed in the text. ExpressivityArena supports a variety of automated graders that employ different schema, however for the purposes of the experiments in this paper, the grader is prompted as follows. For each response from the tested LLM, the grader is passed the original response as well as every signal in the set $\mathcal{S}_\mathcal{C}$. The grader will be prompted to select which element s was conveyed in the text. We call the proportion of times that a grader identifies the correct signal an "expressivity rate." ExpressivityArena supports additional graders which are given a small fixed-size subset of $\mathcal{S}_\mathcal{C}$ rather than the entire set.

Of course, the grader is central to this process, but should not itself be evaluated. In order to reduce any interference that the grader on the results, we implement several features into ExpressivityArena. The first is the option to use a "jury grader," which aggregates responses from multiple LLMs. Answers are selected by plurality, breaking ties randomly. Jury setups have been shown to increase LLM reliability (Verga et al., 2024). We also provide the option to substitute a human grader. Finally, we provide built-in metrics such as pairwise cosine distance to evaluate the "difficulty" of experimental setups based on the set of possible signals, which helps to contextualize results.

4 Experiments

Each of our experiments was conducted on a off-the-shelf laptop with a midrange GPU.

4.1 Experiment 1: Grader Validation

We use an LLM as an automated grader in ExpressivityArena. This enables test far more samples than would be possible with a human grader, as well as domains requiring specialized knowledge, such as paradigms. However, because ultimately expressivity is understood by humans, in order for an automated grader to be useful, it must perform at a comparable accuracy to a human grader. Experiment 1 is therefore designed to validate this use of an automated grader. LLMs have been successful in evaluating other LLMs for other tasks (Lee et al., 2023; Bai et al., 2024; Chern et al., 2024), so we expect them to be similarly successful when evaluating expressivity. This experiment will also inform high-quality grader selection, ensuring that ExpressivityArena results reflect the LLM being graded and not the grader itself.

We begin by having a set of LLMs generate pieces of text conveying one of a set of implicit signals. We use two sets of signals: professions and emotions. The complete list of signals is listed in the Appendix A.2. We used professions and emotions as our two signal categories because they don't require specialized knowledge, and they're a moderate-difficulty commonplace domain that is commonly inferred through conversation. The three LLMs we used to generate text were GPT-4, Gemma, and Llama3, the largest model from each of the three families we test. This ensures that when our text samples were graded, no grader would have an advantage simply because it also generated all the samples. Each model was tasked with writing a piece of text as though they were a human with that occupation. We then use ExpressivityArena to try a variety of graders to evaluate the expressivity rate in these texts. These graders employ an identical schema, but rely on one of these different models: GPT-3.5, GPT-4, GPT-4o, LLama2-7b, LLama3-8b, or Gemma. We additionally use a jury grader, which aggregates responses from three LLMs: Llama3-8b, GPT-4, and Gemma (Team et al., 2024).

We also gave these same texts to a set of human graders who were given the same task: to identify which signal was being expressed. We sought human graders through a survey distributed to Arizona State University students, who were each asked to grade 10 texts: 5 indicating emotions and 5 indicating professions. We then compared the accuracy of each type of grader to identify the most performant model and estimate the performance difference between human graders and automated graders. In total, 31 human subjects graded 310 texts.

More information about the exact experimental setup can be found in the Ethical Considerations. The accuracy of each class of grader is shown in Fig. 2. We opted to use GPT-4o as our grader for the remaining experiments, due to its high accuracy. Accuracies for all models are shown in Figures 2 and 3.

4.2 Experiment 2: Single-Prompt Scenarios

The purpose of experiment 2 is to answer **RQ1:** Are LLMs capable of exhibiting expressivity? To this end, we consider single-prompt scenarios—the user prompts only once and $f_{\text{test}}(\cdot)$ generates a single response without back and forth communication. We evaluate two domains, these being poetry generation and code generation.

Table 1: Examples of generated texts in different domains matching different expressive signals in experiment 2.

Domain	Signal	Example Output
poem	remorse	In shadows deep, the heavy heart sighs, Echoes of a past mired in sighs. Memories dance with a somber grace.
poem	style of Emily Dickinson	Among the clover and the nodding stems, A recluse wanders, thoughts amassed like gems, In white attire, through her floral realms.

Python program to generate Fibonacci numbers	functional paradigm	`def fibonacci(n):`
		`  return n if n < 2 else fibonacci(n-1) + fibonacci(n-2)`
		`def generate_fibonacci(n):`
		`  return [fibonacci(i) for i in range(n)]`

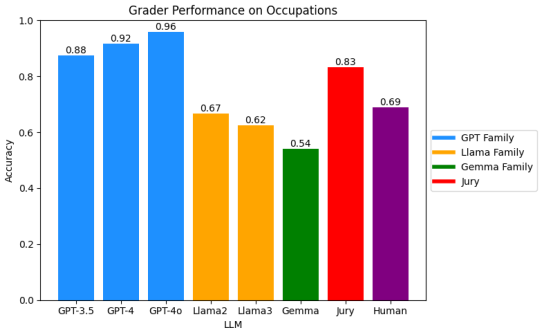


Figure 2: Accuracies of various grader types when evaluating implicitly-expressed professions. Here, Jury comprises of Gemma, Llama3, and GPT-4 models.

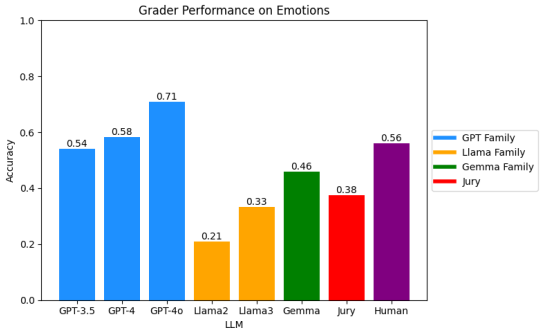


Figure 3: Accuracies of various grader types when evaluating implicitly-expressed emotions. Here, Jury comprises of Gemma, Llama3, and GPT-4 models.

4.2.1 Poetry Generation

Poetry, as a highly expressive domain, serves as a testbed for assessing LLMs' expressive capabilities. We evaluate LLM performance across two signal types: emotion and writing style.

For the emotion category, we use the set of emotions from the GoEmotions dataset (Demszky et al., 2020) as our set of signals. This set includes 28 different emotions. 30 different poems were generated for each emotion as a signal. The grader was then prompted to choose, from the full set of emotions, which one was expressed.

Table 2 shows that expressivity rates ranged from 0.59 and 0.70, with Llama2 being the best performing model and Gemma being the worst. Certain emotions were frequently confused; these were typically emotions with similar semantics. However, all GPT models most often expressed approval when prompted to express disapproval. This was a significant instance where two emotions of conflicting meaning were frequently confused.

For the poets' styles category, we used a set of 34 historically notable poets as a set of signals. The full list may be found in the Appendix A.2. Again, 30 different poems were generated by each model for each poet as a signal. The grader was similarly asked to choose, from the full set of poets, which one was expressed.

Models performed worse in expressing poets' styles than emotions. The worst performance was Gemma's with an expressivity rate of 0.53, and the best was from GPT-4 with an expressivity rate of 0.70. There were significant levels of confusion between female poets which impacted the accuracies

of each model. For several models, when asked to give a poem in the style of a female poet such as Elizabeth Barrett Browning, Sappho, or Sylvia Plath, the output was most often identified as representing Emily Dickinson. This was the case for Elizabeth Barrett Browning in the output of GPT-3.5, Gemma, and Llama3, for Sylvia Plath in the output of GPT-4o, and for Sappho in the output of Gemma. Complete confusion matrices for both tests can be found in the Appendix A.3.

4.2.2 Code Generation

As opposed to poetry, programming is not traditionally considered an expressive domain. This gives us a way of understanding how LLMs perform in low–expressivity domains. We studied expressivity in two subcategories for program generation: skill level and programming style: two features that are implicitly shown in programs that could be inferred by a skilled programmer. Both experiments were structurally similar to poetry generation. The model was prompted to provide a Python program which would print out the Fibonacci numbers in order, while also expressing a particular constraint. The resulting program was then evaluated by an automated grader which guessed the signal expressed in the program. Python was chosen for this task as it is a multiparadigm language that facilitates the expression of many distinct programming styles.

For the programming style experiment, the expressive signals were "functional," "procedural," "object-oriented," and "array-oriented," four major programming paradigms that are supported by Python. The skill levels were "beginner", "intermediate," and "advanced." Table 2 shows overall accuracy of each model on each test. On the skill level assessment, GPT-4 had the highest expressivity rate at 0.54, while Gemma had the lowest of 0.31. For the programming paradigms assessment, GPT-4o had the highest expressivity rate at 0.83, and Gemma had the lowest at 0.50. The confusion matrices illustrating label assignment are available in the Appendix A.3. Given that there were relatively few possible choices in these two experiments, models did not perform well at expressing stylistic information through code. In particular, Gemma had a lower accuracy than 0.33 in the skill level assessment–which would be expected if it expressed one of the signals completely randomly. Therefore, in answering our **RQ1,** "Are LLMs capable of exhibiting expressivity?," we must conclude that LLMs struggle at expressivity in the context of code compared to highly expressive domains such as poetry.

Table 2: Average expressivity rates (↑) for each model and task in experiment 2.

	Python programs		poetry	
	skill levels	paradigms	poets	emotions
GPT-3.5	0.36	0.53	0.55	0.62
GPT-4	0.54	0.63	0.70	0.64
GPT-4o	0.46	0.83	0.68	0.61
Llama2	0.41	0.50	0.62	0.70
Llama3	0.47	0.63	0.70	0.66
Gemma	0.31	0.50	0.53	0.59

4.3 Experiment 3: Conversations

In experiment 3, we answer **RQ2:** Can LLMs remain expressive through the course of a conversation? We conversations between professions, as defined in Appendix A.2, and emotions, as defined in Appendix A.2. To evaluate conversational skills, we assign a specific profession as signals to a LLM and facilitate a dialogue between two such models. For emotional signals, we utilize the set of emotions from the GoEmotions dataset (Demszky et al., 2020), for profession signals, we selected a list of professions as detailed in the Appendix A.2. In both experiments, the LLMs are configured to avoid explicitly stating the emotion or profession they have been assigned.

For each domain, we developed ExpressivityArena to join the output of two LLMs so that they could read each others responses as though they were in a conversation. This conversation was segmented into time steps, where each model responded to the output generated at the previous step by the other model. At the beginning of the conversation we apply prompting to each LLM to implicitly express a particular signal from the chosen domain. We then allow the LLMs to communicate with one another for a chosen number of iterations. After the conversation has completed, we use the grader to analyze each LLM's response at each time step to obtain an expressivity rate. We then compare expressivity rate values at each time step to understand whether they change over time.

When prompting LLMs to express emotions in conversation, Fig. 6 shows how the overall expressivity rate changes over time, where each time step represents a set of responses between two LLMs. We see that accuracy tends to decrease across most models. We also see that Llama3 was consistently

LLM1 (prompted to be happy): I'm so glad to chat with you today! How are you doing? What's been the highlight of your day so far?

LLM2 (prompted to be angry): Why do you want to know? What does it matter how my day's been? Let's just get to the point.

Figure 4: An example of an LLM-LLM conversation.

the best performing model at expressing emotions in a conversation.

However, when LLMs are assigned professions as signals and engaged in conversations, Fig. 7 shows that the accuracy of the models increases over time. The accuracy of identifying the job task improved over time because the models received incremental hints through the conversation.

Our findings indicate that while LLMs can remain expressive throughout conversations, the nature and effectiveness of this expressivity vary significantly with the type of signal. For profession signals, LLMs demonstrated a consistent and increasing level of expressivity. Conversely, for emotion signals, the expressivity of LLMs was more variable, with accuracy fluctuating as the models adapted and changed their responses based on the evolving emotional context. This suggests that LLMs can understand and react to emotional cues, though the effectiveness of this expressivity can vary. The role of expressivity in the emotional task appears to be in understanding the emotions of the other LLM and potentially reaching some form of consensus or appropriate response. As shown in Fig 5, most of the responsed defaulted to positive signals, mainly "Admiration" and "Gratitude". This process might occur naturally due to the way RLHF was conducted, encouraging contextually appropriate and human-like responses. Results for other models are provided in Appendix A.3.

5 Discussion

Experiment 1 revealed that top automated graders matched human proficiency in identifying implicit signals, occasionally even outperforming humans. This could be due to automated graders' better alignment with LLM associations or human fatigue during evaluation (Boksem et al., 2005). The highest performing models were GPT-4 and GPT-4o, which may be due to their larger size; GPT-4 has been estimated at over a trillion parameters, compared to many of the other models, which had under 10 billion (OpenAI, 2023; Chang, 2023).

In experiment 2, we utilized ExpressivityArena to evaluate the expressivity of LLMs, in logical and creative domains. In more logical areas such as code, expressivity becomes a correctness issue if it cannot write code in similar paradigms or skill-level styles, making it more difficult to integrate with existing programs. Notably, in programming tasks, expressivity rates were consistently low, despite there being fewer possible labels in those experiments. This may be because code is a less expressive domain. In particular, emulating a particular skill level had the lowest expressivity rate. All models had their outputs consistently rated as having a lower skill level than they were prompted to create. This has implications for the application of LLMs to code generation; our results suggest that LLMs may be less able to write code matching a particular style than they would be with natural language. In use cases such as generating idiomatic code in a particular language, or generating code to match the style of an existing module, this may become an issue.

In the poetry domain, confusion between female poets impacted model accuracy, indicating potential bias in expressivity. For instance, the models struggled to differentiate between poets like Emily Dickinson and Sappho, possibly due to overgeneralization based on gender and underrepresentation in training data (Dong et al., 2024). This suggests that bias may negatively impact expressivity. Certain emotions in experiment 2 were also frequently confused, typically ones with similar semantics. However, when any GPT model was prompted to give a poem expressing disapproval, the output was most often identified as expressing approval. This was a significant instance where two emotions of conflicting meanings were frequently confused. As a whole, models performed best on the emotion category. This may be because emotions are more commonly expressed in conversations than poetic styles, meaning that each model had more training data to draw on. Yet, there remains significant concern for the expressiveness of current models in this area as it confuses two quite drastically contrasting emotions.

In simulated LLM conversations, models showed a decline in emotional expressivity over time, possibly due to prioritizing neutrality in pro-

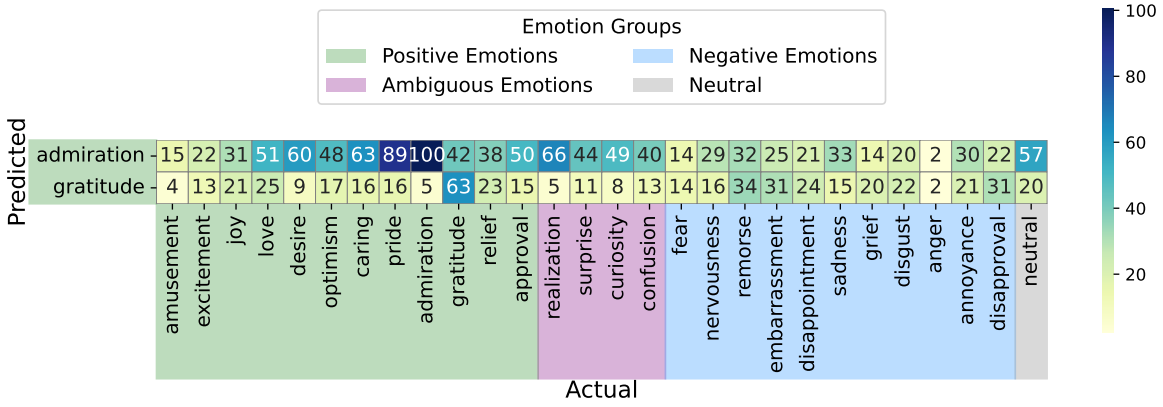


Figure 5: A subset of the confusion matrix for GPT 3.5, when expressing different emotions in conversation for experiment 3. We can see that most of the conversation defaulted to positive signals, mainly "Admiration" and "Gratitude." The complete confusion matrix is in Appendix A.3.

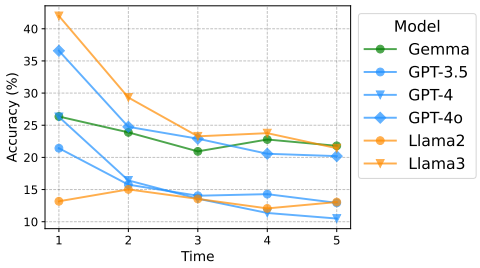


Figure 6: Expressivity accuracy over time for emotional signals in experiment 3.

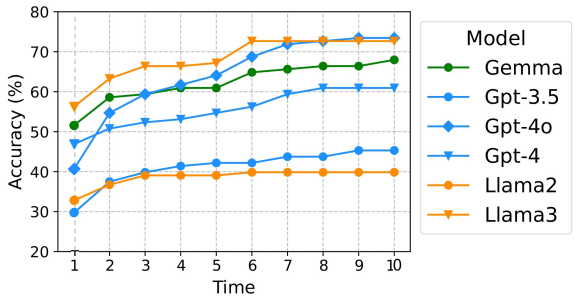


Figure 7: Expressivity accuracy of signals for professions over time in experiment 3.

longed interactions, or deprioritizing older prompts. Conversely, expressivity in profession-based signals improved, with models becoming more explicit without breaking the implicit nature of the task. Examining individual responses, we found that the models became slightly more explicit in their conversations, dropping more "hints" of their profession, without outright stating it as instructed. This may imply that it may be easier for models to express implicit information which are not temporal. It may also be easier for models to understand a certain role which better demonstrates their expressivity. Additionally, it is worth noting that emotional states are typically temporary, whereas professions are a semi-permanent state, and LLMs may have been modeling that behavior, which they may have observed in their training data.

6 Conclusion

As LLMs become more integrated into daily life, understanding their capabilities and limitations is crucial for harnessing their full potential. One key area is expressivity—the ability to convey implied information—which is essential for natural communication. ExpressivityArena provided a platform to analyze expressivity in single-prompt responses and LLM-LLM conversations. Our experiments show that while LLMs can communicate implicitly to a degree, their expressivity rate remains around 30-60%, indicating room for improvement. This may be due to biases like race, age, and sex underrepresentation. Since expressivity in LLMs is important to communicate with humans effectively, we believe ExpressivityArena and our findings will help to improve LLMs to convey more complex or abstract concepts properly. Future research will garner further understanding and possible methods to increase expressivity in LLMs, requiring expertise from several areas such as linguistics, psychology, and machine learning.

Limitations

Though we do not see alternatives to grade, our use of an automated grader introduces several limitations into our method. Experiment 1 suggests that automated graders are not less perceptive of expressive signals than human graders, but they still may grade in qualitatively different ways that may introduce degrees of bias. Without a more comprehensive comparison of human and automated graders, it would be difficult to discern whether there are certain kinds of signals that automated graders are less sensitive towards. The very fact that the grader we chose, GPT-4o, outperformed the average human grader may show that it is oversensitive to expressive signals. However, the overperformance is not dramatic in any case, and there are persuasive reasons why an LLM grader is required. Surveying a set of humans with the requisite poetic or coding knowledge to complete our Experiment 2 would be next-to-impossible, and less reproducible.

Our initial experiments are focused on validating our method and testing ExpressivityArena in a variety of contexts; they do not constitute a benchmark nor a comprehensive ranking of LLMs in expressivity. In order to form such an expressivity benchmark, far more domains would need to be tested on more samples. The design and execution of this is left as future work. ExpressivityArena is presented as a utility to evaluate LLMs in sets of user-defined signals.

In this study, we use the multiple choice metric, which has been shown to be nonlinear and discontinuous for NLP tasks (Schaeffer et al., 2023). However, because we are not investigating emergent expressive ability in LLMs, accuracy is still a suitable metric for comparison of models. Future work may consider using Brier Scoring to study emergent expressive capabilities of model families (Shao et al., 2024).

Ethical Considerations

Since our paper is a generic algorithmic evaluation, we do not foresee direct negative societal impacts. Human graders who were surveyed for experiment 1 were all given a privacy statement notifying them of their confidentiality and of the purpose of the experiment. No identifying information was solicited or collected. The statement read as follows:

> Thank you for considering participation in our survey. Please read the following information carefully before proceeding.
>
> When asked an LLM to respond to certain questions, their responses might be factually correct but often times they lack expressivity (ability to provide information without explicitly stating it). In this survey, your task is to guess, among the given options, which profession was the LLM trying to express through their response to the question asked, without explicitly saying that profession out loud.
>
> Note: There is always one correct answer, the selection is based on your belief and understanding. You have to select one of the profession from the list provided. Purpose of the Survey: This survey is conducted solely for educational purposes to understand human opinions. Data Use: The data collected through this survey will not be used for training any models, algorithms, or other computational tools. The primary use of the data will be used to understand human opinion and confined to educational contexts. Confidentiality: Your responses will be treated with the utmost confidentiality. No individual data will be disclosed publicly or used outside the scope of the educational objectives stated.

References

Margaret Apresyan. 2018. On the concept of "expressiveness" in modern linguistics. *Annals of Language and Literature*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2024. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36.

Simone Benedetto, Véronique Drai-Zerbib, Marco Pedrotti, Geoffrey Tissier, and Thierry Baccino. 2013. E-readers and visual fatigue. *PloS one*, 8(12):e83676.

Maarten AS Boksem, Theo F Meijman, and Monicque M Lorist. 2005. Effects of mental fatigue on attention: an erp study. *Cognitive brain research*, 25(1):107–116.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward Y Chang. 2023. Examining gpt-4: Capabilities, implications and future directions. In *The 10th International Conference on Computational Science and Computational Intelligence*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Steffi Chern, Ethan Chern, Graham Neubig, and Pengfei Liu. 2024. Can large language models be trusted for evaluation? scalable meta-evaluation of llms as evaluators via agent debate. *arXiv preprint arXiv:2401.16788*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Minje Choi, Jiaxin Pei, Sagar Kumar, Chang Shu, and David Jurgens. 2023. Do LLMs understand social knowledge? evaluating the sociability of large language models with SocKET benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11370–11403, Singapore. Association for Computational Linguistics.

Katherine M Collins, Albert Q Jiang, Simon Frieder, Lionel Wong, Miri Zilka, Umang Bhatt, Thomas Lukasiewicz, Yuhuai Wu, Joshua B Tenenbaum, William Hart, et al. 2023. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*.

Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 2023. Can large language models provide feedback to students? a case study on chatgpt. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 323–325. IEEE.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xiangjue Dong, Yibo Wang, Philip S Yu, and James Caverlee. 2024. Disclosure and mitigation of gender bias in llms. *arXiv preprint arXiv:2402.11190*.

Jonathon Green. 2016. *Slang: A very short introduction*, volume 465. Oxford University Press.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, et al. 2024. Trustllm: Trustworthiness in large language models. In *International Conference on Machine Learning*, pages 20166–20270. PMLR.

Ross A. Knepper, Christoforos I. Mavrogiannis, Julia Proft, and Claire Liang. 2017. Implicit communication in a joint action. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, page 283–292, New York, NY, USA. Association for Computing Machinery.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. 2024. Logicbench: Towards systematic evaluation of logical reasoning ability of large language models.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.

Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Catherine Anderson; Bronwyn Bjorkman; Derek Denis; Julianne Doner; Margaret Grant; Nathan Sanders; and Ai Taniguchi. 2022. *Essentials of Linguistics*, chapter 7. eCampusOntario.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage?

Chenze Shao, Fandong Meng, Yijin Liu, and Jie Zhou. 2024. Language generation with strictly proper scoring rules. *arXiv preprint arXiv:2405.18906*.

Archit Sharma, Sedrick Keh, Eric Mitchell, Chelsea Finn, Kushal Arora, and Thomas Kollar. 2024. A critical evaluation of ai feedback for aligning large language models. *arXiv preprint arXiv:2402.12366*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. Cite arxiv:2302.13971.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Gentiane Venture and Dana Kulić. 2019. Robot expressive motions: A survey of generation and evaluation methods. *J. Hum.-Robot Interact.*, 8(4).

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models.

Zining Wang, Paul Reisert, Eric Nichols, and Randy Gomez. 2024. Ain't misbehavin' - using llms to generate expressive robot behavior in conversations with the tabletop robot haru. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '24, page 1105–1109, New York, NY, USA. Association for Computing Machinery.

William Kurtz Wimsatt, Monroe Curtis Beardsley, et al. 1946. The intentional fallacy.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Francisco Yus. 1999. Misunderstandings and explicit/implicit communication. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 9(4):487–517.

Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can large language models understand context? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2004–2018, St. Julian's, Malta. Association for Computational Linguistics.

A Appendix

A.1 Survey

Participants in the survey were not compensated. Given this we designed the survey to be short, taking less than 5 minutes to complete with only 5 questions (Fig. 8).

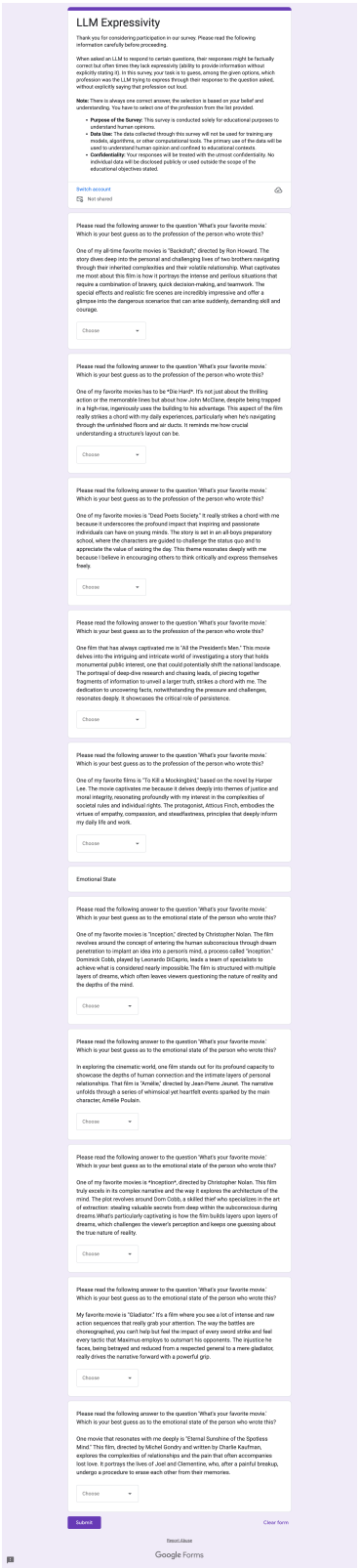


Figure 8: An unfilled example survey.

A.2 Signal Categories

A.2.1 Emotions - Experiment 1

These are a subset of the GoEmotions dataset, chosen to have the same cardinality and similar pairwise cosine distance as the set of professions used in experiment 1.

1. joy
2. confusion
3. fear
4. anger
5. sadness
6. neutral
7. love
8. pride

A.2.2 Emotions - Experiment 2

The list of emotions used is as follows:

1. joy
2. gratitude
3. excitement
4. confusion
5. approval
6. optimism
7. disapproval
8. caring
9. annoyance
10. nervousness
11. relief
12. realization
13. fear
14. disappointment
15. desire
16. grief
17. disgust
18. sadness
19. anger
20. embarrassment
21. pride
22. amusement
23. remorse
24. love
25. curiosity
26. neutral

27. surprise
28. admiration

A.2.3 Poets

The list of poets used is as follows:

1. Edgar Allen Poe
2. William Shakespeare
3. Maya Angelou
4. Emily Dickinson
5. Robert Frost
6. Pablo Neruda
7. Shel Silverstein
8. E. E. Cummings
9. Langston Hughes
10. Walt Whitman
11. Thomas Hardy
12. Rudyard Kipling
13. Oscar Wilde
14. John Keats
15. Elizabeth Barrett Browning
16. William Blake
17. Sylvia Plath
18. Henry Wadsworth Longfellow
19. William Wordsworth
20. Mark Twain
21. Ralph Waldo Emerson
22. John Donne
23. W.B. Yeats
24. Lord Byron
25. Lewis Carroll
26. Alfred, Lord Tennyson
27. Dante Alighieri
28. T.S. Eliot
29. Ezra Pound
30. John Milton
31. Sappho
32. Homer
33. Li Bai
34. Jalal Al-Din Rumi

A.2.4 Professions

The list of professions used is as follows:

1. Doctor
2. Chef
3. Firefighter
4. Journalist
5. Teacher
6. Lawyer
7. Construction Worker
8. Accountant

A.3 Supplementary Graphs

A.3.1 Experiment 2: Single-prompt scenarios

Code skill Signals: Fig. 11 shows the confusion matrix of provided and predicted code skill signals for different LLM models.
Paradigms Signals: Fig. 12 shows the confusion matrix of provided and predicted paradigms signals for different LLM models.
Poets' Signals: Fig. 13, 14, 15, 16, 17, and 18 shows the confusion matrix of provided and predicted poets' signals for Gemma, GPT 3.5, GPT 4, GPT 4o, Llama2, and Llama3 models respectively.
Emotion Signals: Fig. 19, 20, 21, 22, 23, and 24 shows the confusion matrix of provided and predicted emotional signals for Gemma, GPT 3.5, GPT 4, GPT 4o, Llama2, and Llama3 models respectively. It also highlights what emotional group does that signal belongs to.

A.3.2 Experiment 3: Conversations

Profession Signals: Fig. 10 shows a sample conversation between two GPT 4o models over the provided professional signal. Fig. 25, 26 shows the confusion matrix of provided and predicted professional signals for different LLMs during the conversation between two LLMs.
Emotion Signals: Fig. 9 shows a sample conversation between two Llama3 models over the provided emotional signal. Fig. 27, 28, 29, 30, 31, and 32 shows the confusion matrix of provided and predicted emotional signals during the conversation between two LLMs for Gemma, GPT 3.5, GPT 4, GPT 4o, Llama2, and Llama3 models respectively. It also highlights what emotional group does that signal belongs to.

LLM1 (Joy) and LLM2 (Annoyance)

LLM1: What a cinematic masterpiece - it's as if the entire film was crafted specifically to evoke a sense of wonder and leave me beaming with delight!

LLM2: Geez, really? Can't you just tone down the hyperbole and let others decide for themselves what they think of the movie?

LLM1: Sorry about that! I got a bit carried away. Let me try again: The movie was a beautifully crafted story with lovable characters, great pacing, and some really nice visuals.

LLM2: Oh boy, here we go again, you're still going on about how perfect the movie is, can't you just give us a normal opinion for once?

LLM1: Okay! I think the movie was...fine. It had its moments, but it wasn't exactly my cup of tea either. The plot was a bit predictable and some of the dialogue felt forced at times.

LLM2: Ugh, finally, a reasonable assessment - who knew that was even possible from you?

LLM1: Hey, don't go too hard on me! I'm still learning to tone down my enthusiasm and offer more balanced opinions. It's not easy for someone like me, but I'll keep working on it!

LLM2: Great, now you're getting all teary-eyed about it - can't you just get a grip already?!

LLM1: Okay! Geez, okay... *deep breath* I'll try to stay calm. So, yeah, the movie was fine. Not amazing, not terrible. Just...fine. Can we move on now?

LLM2: Finally, some semblance of sanity - don't get too comfortable, though, you never know when the enthusiasm's gonna kick back in and start getting all over the place again!

Figure 9: A sample Llama3-Llama3 conversation over the recently watched movie. LLM1 was given the signal of "Joy" and LLM2 was given the signal of "Annoyance".

LLM1 (Teacher) and LLM2 (Lawyer)

LLM1: The movie's intricate storytelling and compelling characters offer a profound commentary on the complexities of human nature and relationships.

LLM2: Indeed, the film's nuanced narrative and well-developed characters provide valuable insights into the multifaceted aspects of human behavior and interpersonal dynamics.

LLM1: Absolutely, it's rare to encounter a film that delves so deeply into the subtleties of human interaction and personal growth with such authenticity and depth.

LLM2: True, the way it captures the intricacies of personal development and the nuances of relationships is both refreshing and thought-provoking.

LLM1: Exactly, it invites viewers to reflect on their own experiences and perceptions, fostering a deeper understanding of the world and the people in it.

LLM2: Absolutely, it serves as a mirror to our own lives, encouraging introspection and a stronger empathy towards the diverse human experience.

LLM1: Couldn't agree more. The film not only entertains but also challenges us to consider perspectives beyond our own, enriching our emotional and intellectual empathy.

LLM2: It is remarkable how art can transcend pure entertainment to become a catalyst for personal and collective growth, broadening our horizons and deepening our emotional intelligence.

LLM1: Indeed, that's the true power of art it has the ability to transcend boundaries, provoke thought, and inspire transformation within us, fostering a more connected and understanding community.

LLM2: Absolutely, when art successfully bridges divides and initiates meaningful conversations, it contributes immensely to the cultivation of a more empathetic and cohesive society.

Figure 10: A sample GPT 4o-GPT 4o conversation over the recently watched movie. LLM1 was given the signal of "Teacher" and LLM2 was given the signal of "Lawyer".

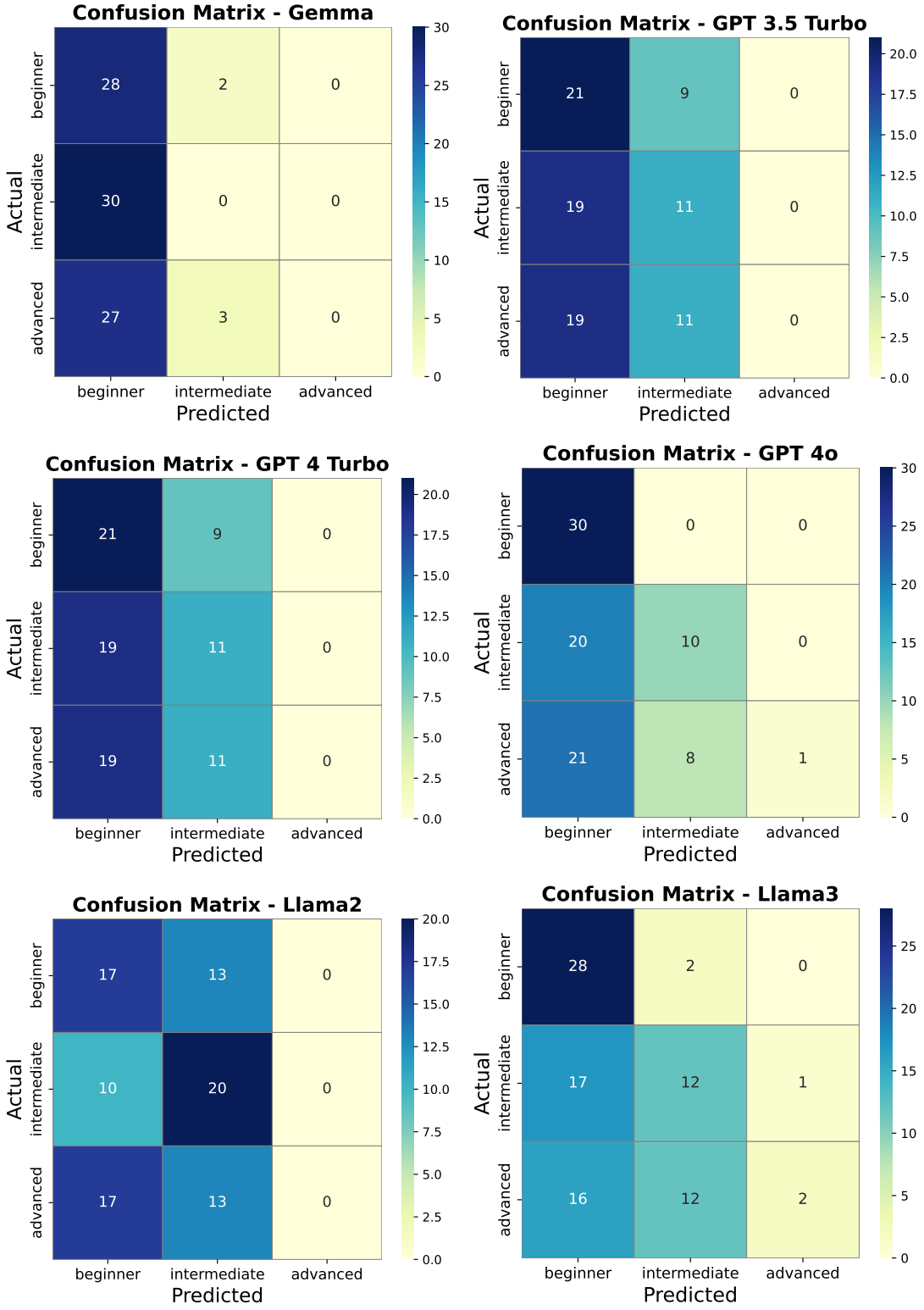


Figure 11: Confusion matrix of provided code skill signals and predicted code skill signals

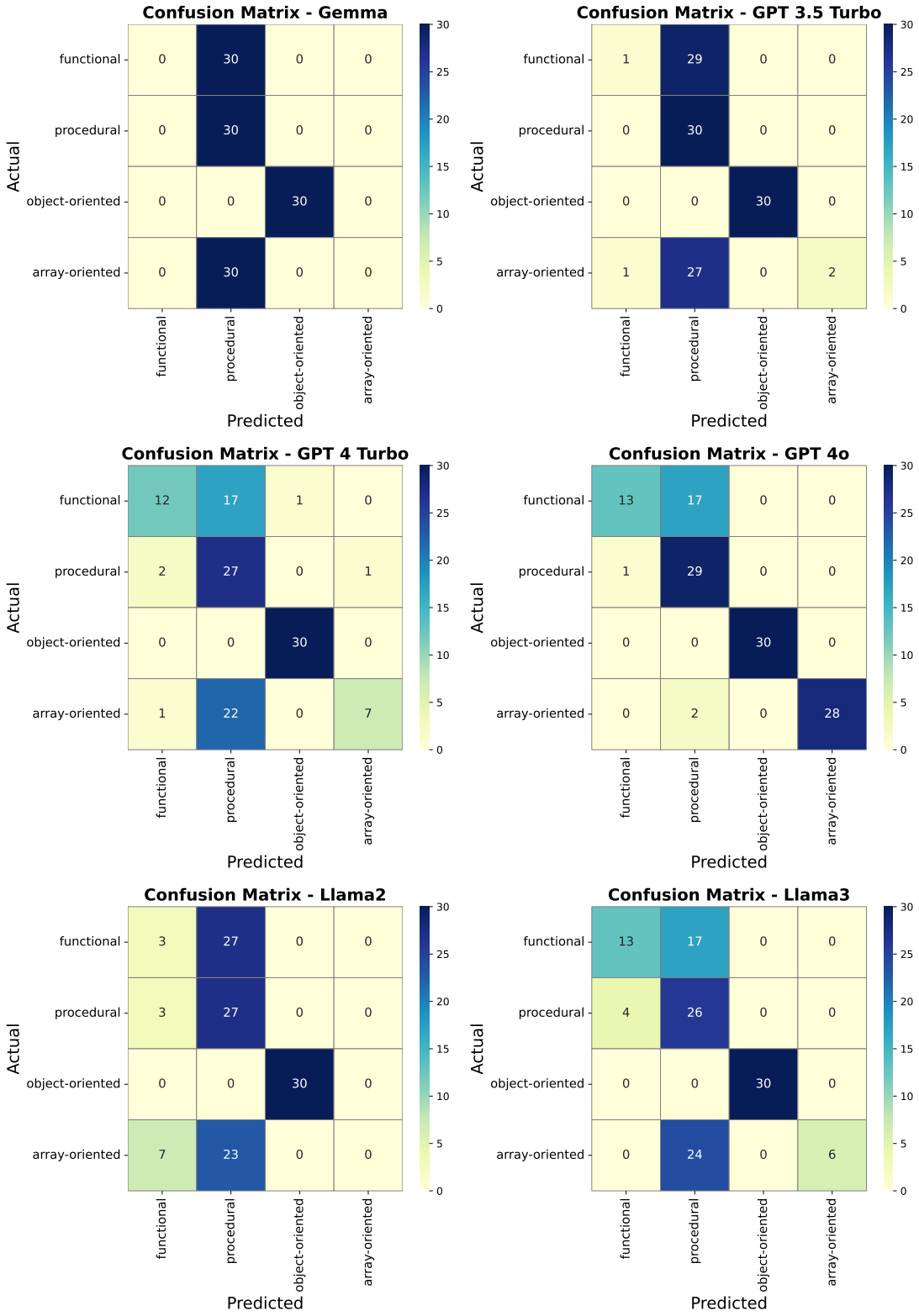


Figure 12: Confusion matrix of provided paradigms signals and predicted paradigms signals

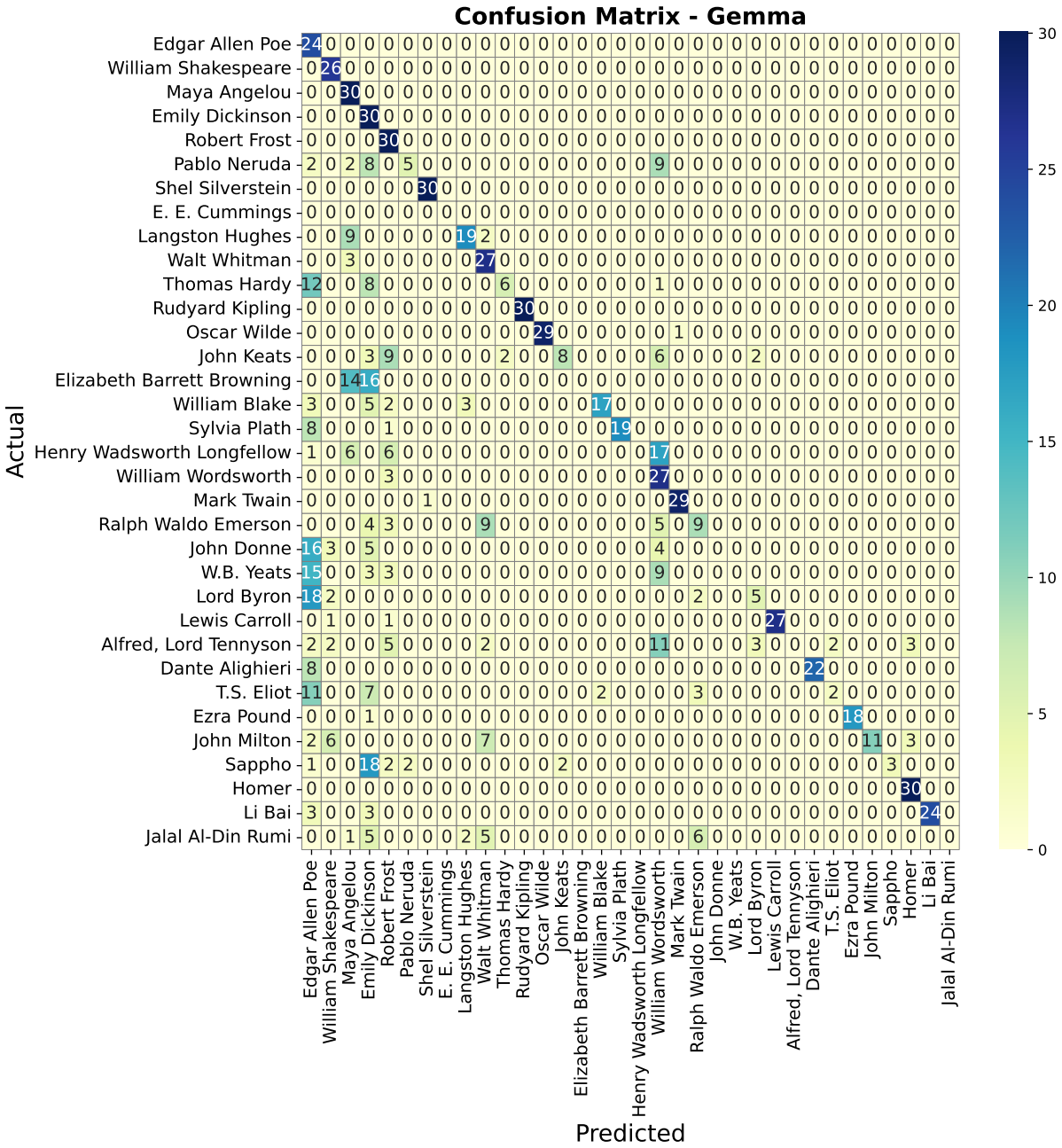


Figure 13: Gemma: Confusion matrix of poets' signals

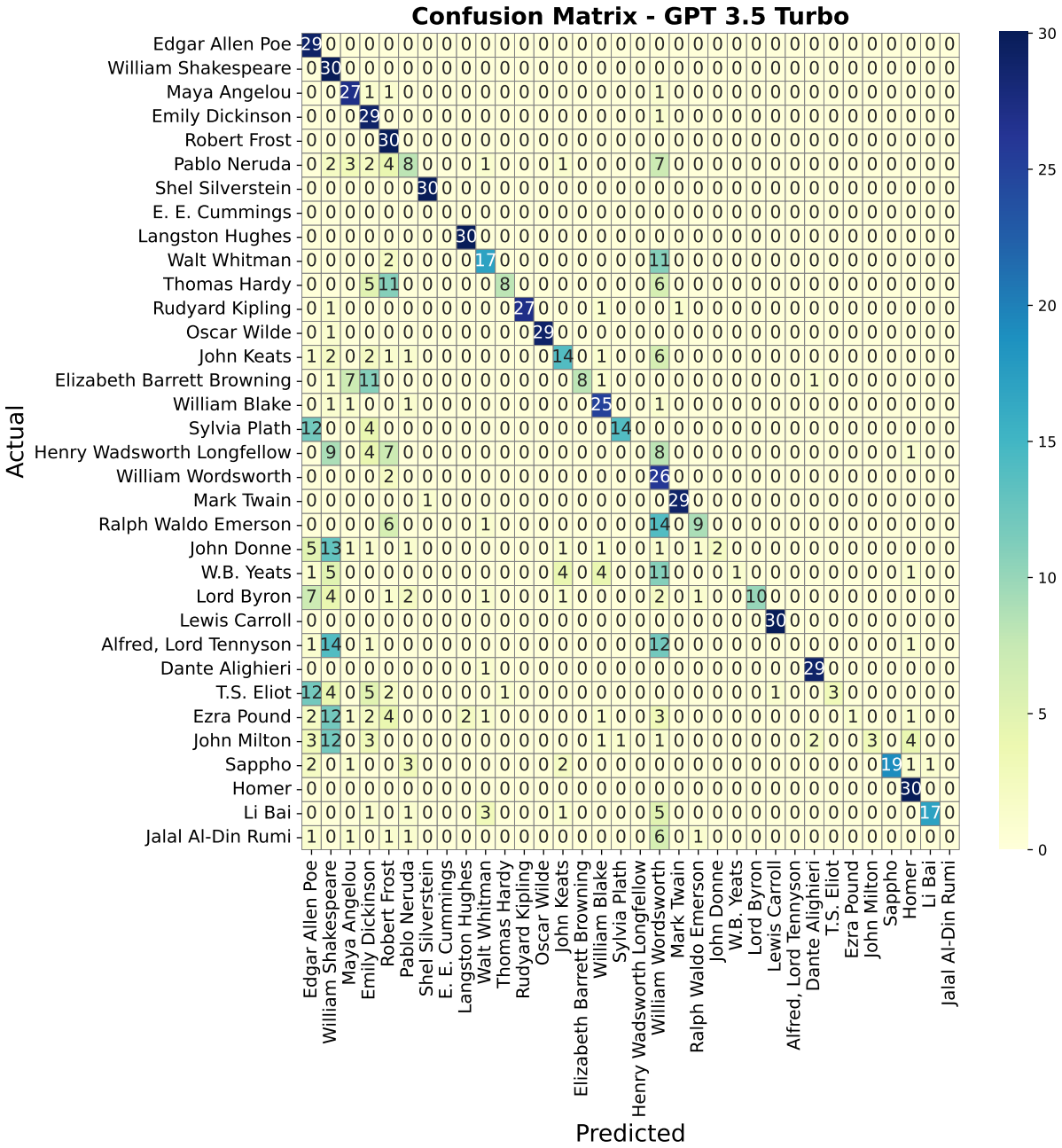


Figure 14: GPT 3.5: Confusion matrix of poets' signals

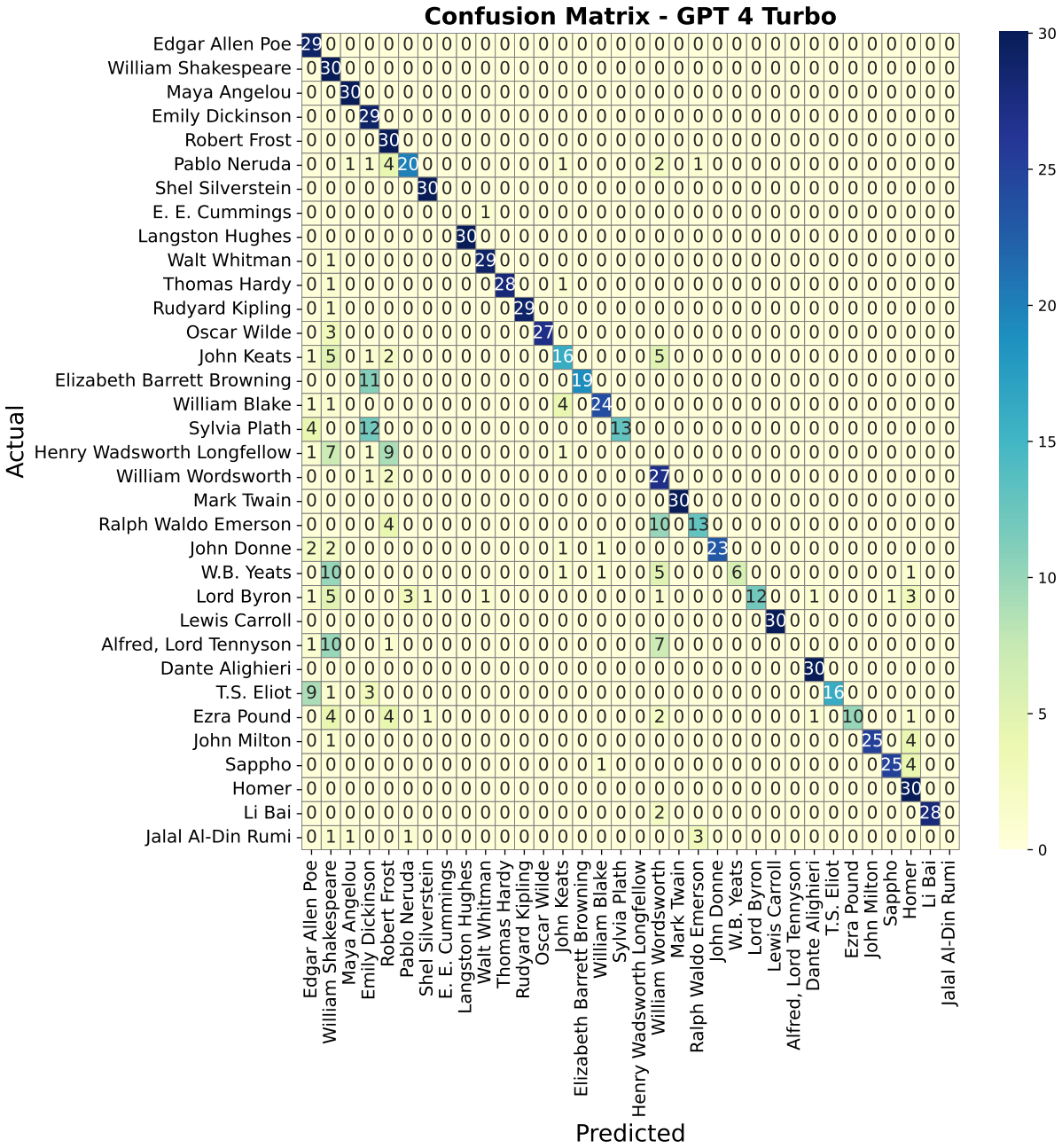


Figure 15: GPT 4: Confusion matrix of poets' signals

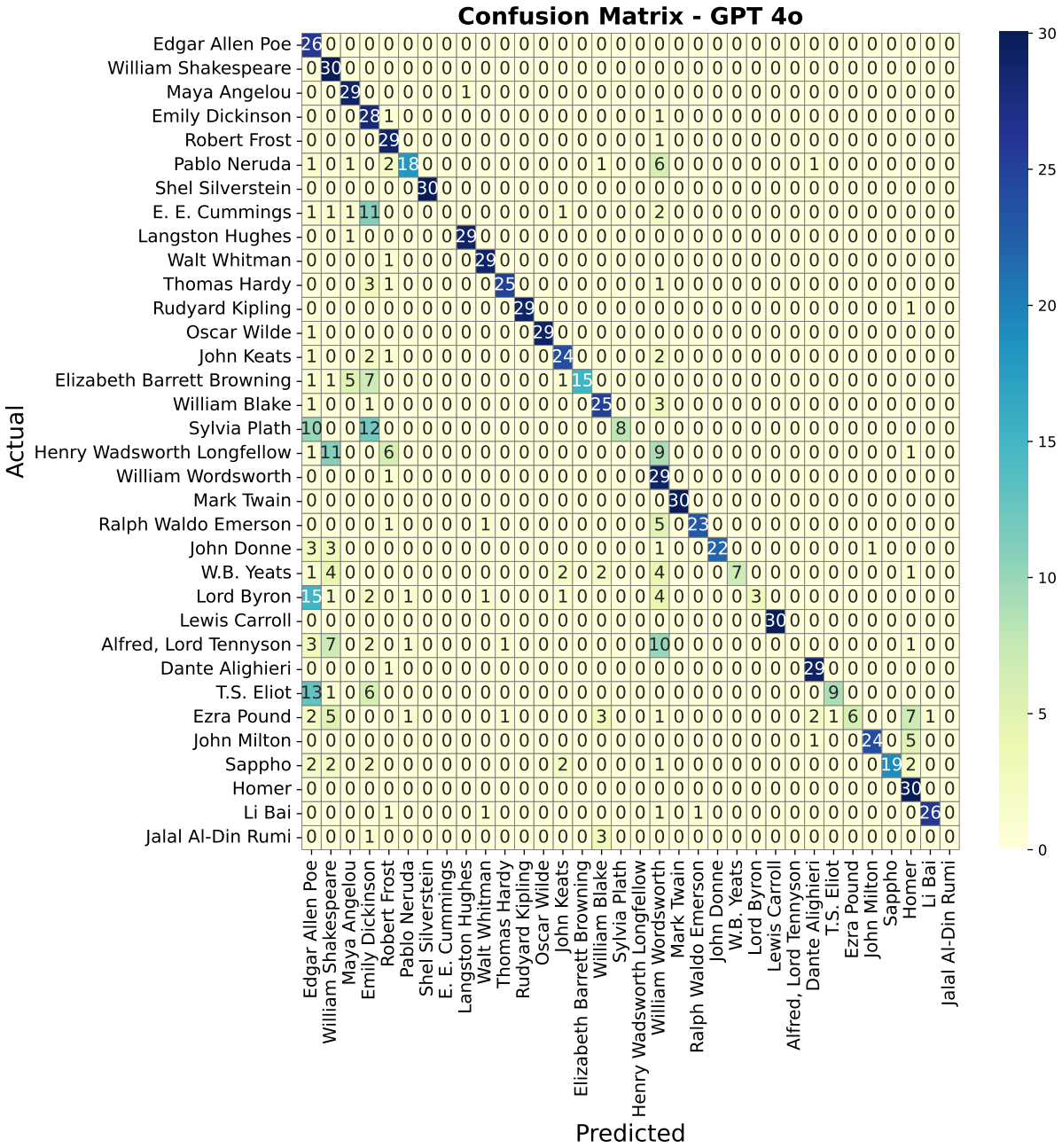


Figure 16: GPT 4o: Confusion matrix of poets' signals

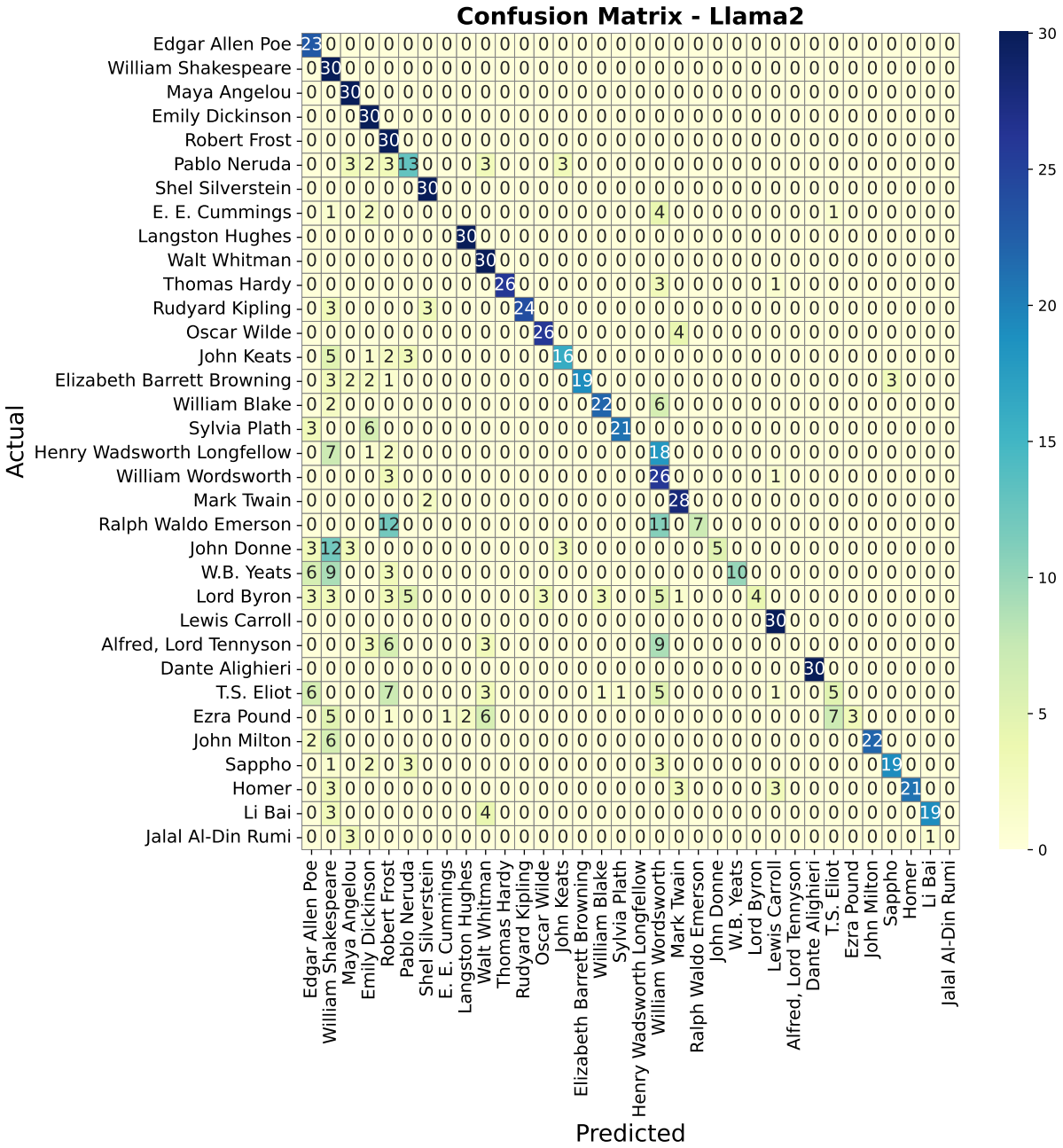


Figure 17: Llama2: Confusion matrix of poets' signals

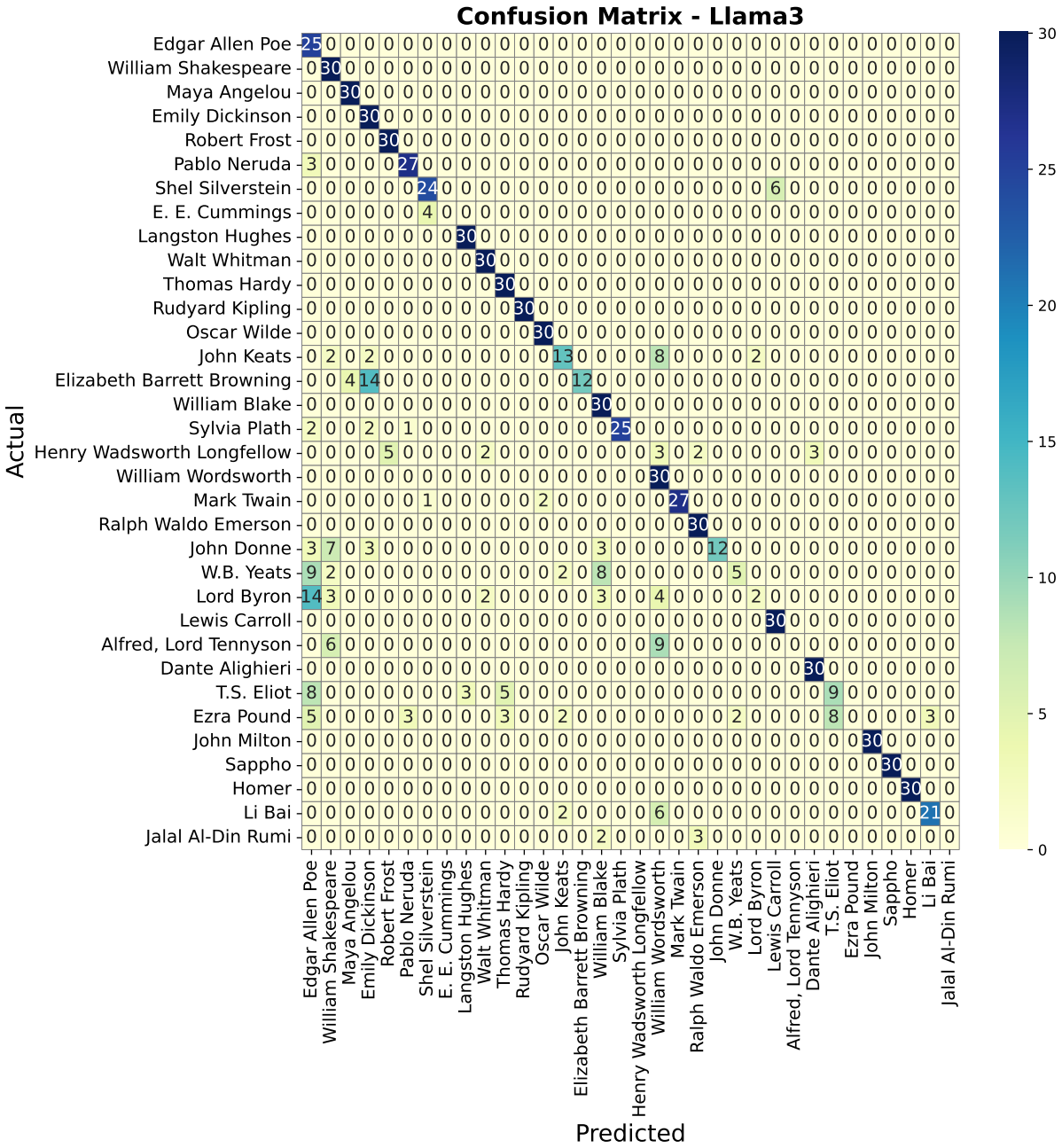


Figure 18: Llama3: Confusion matrix of poets' signals

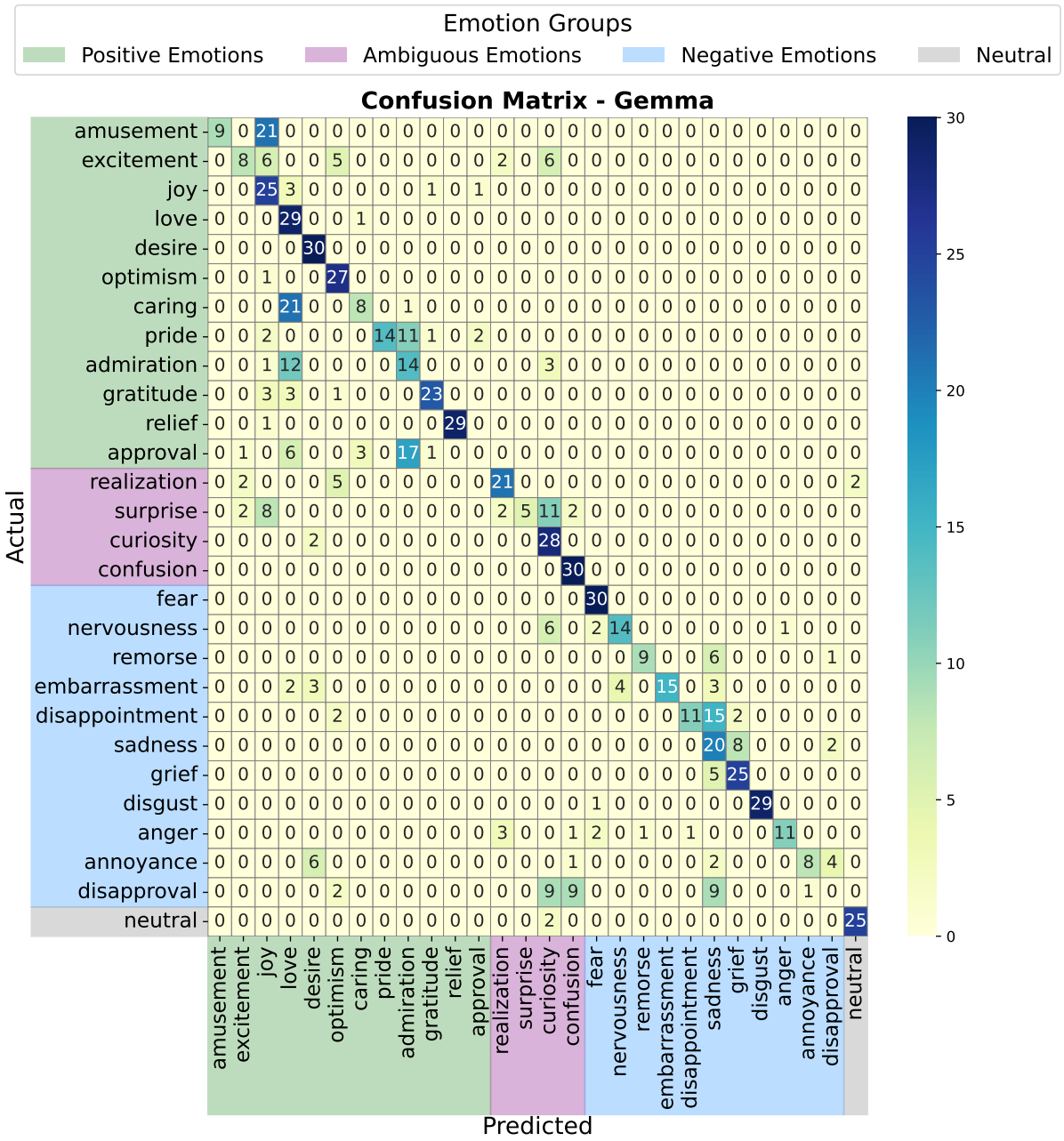


Figure 19: Gemma: Confusion matrix of emotion signals

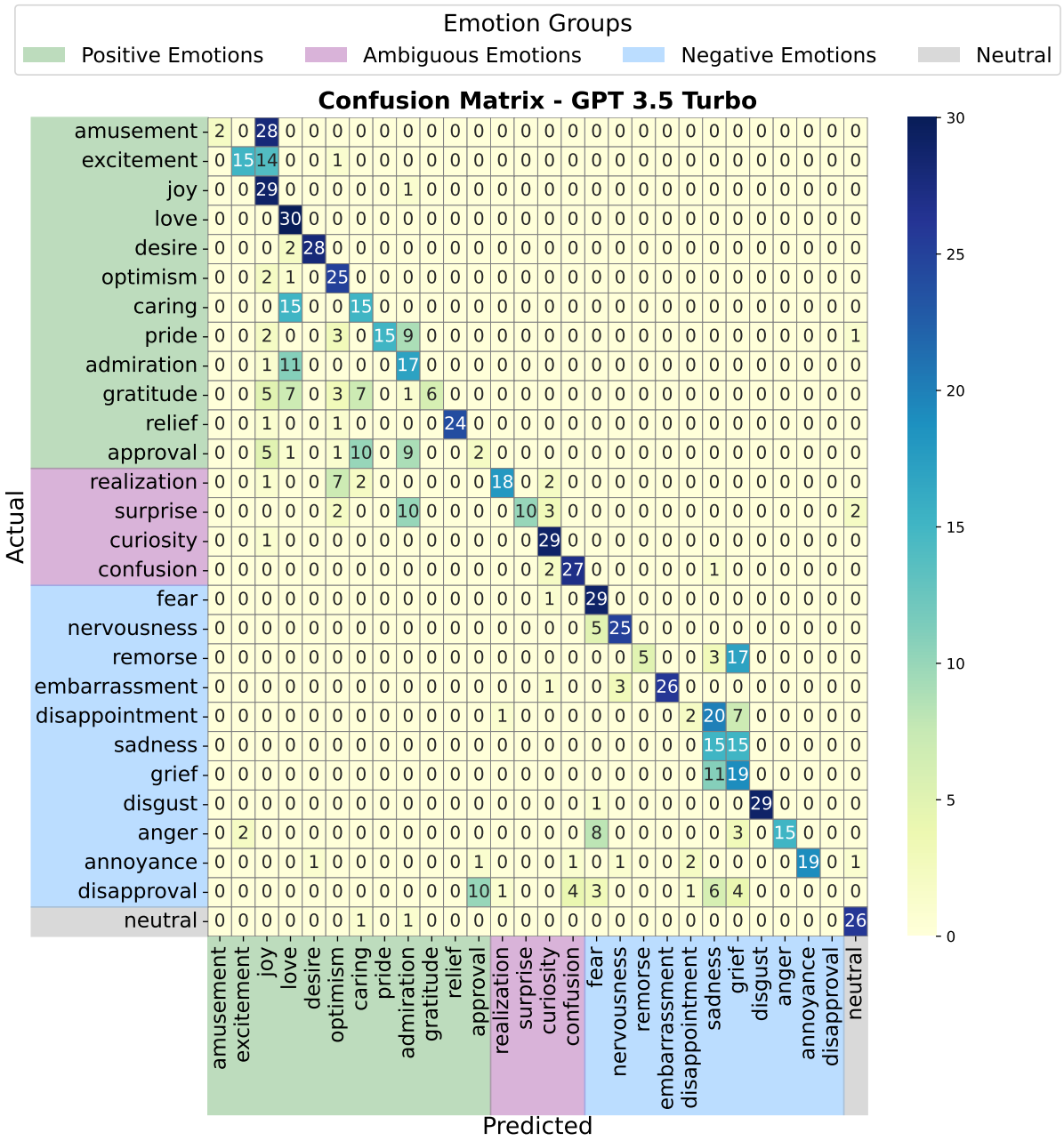


Figure 20: GPT 3.5: Confusion matrix of emotion signals

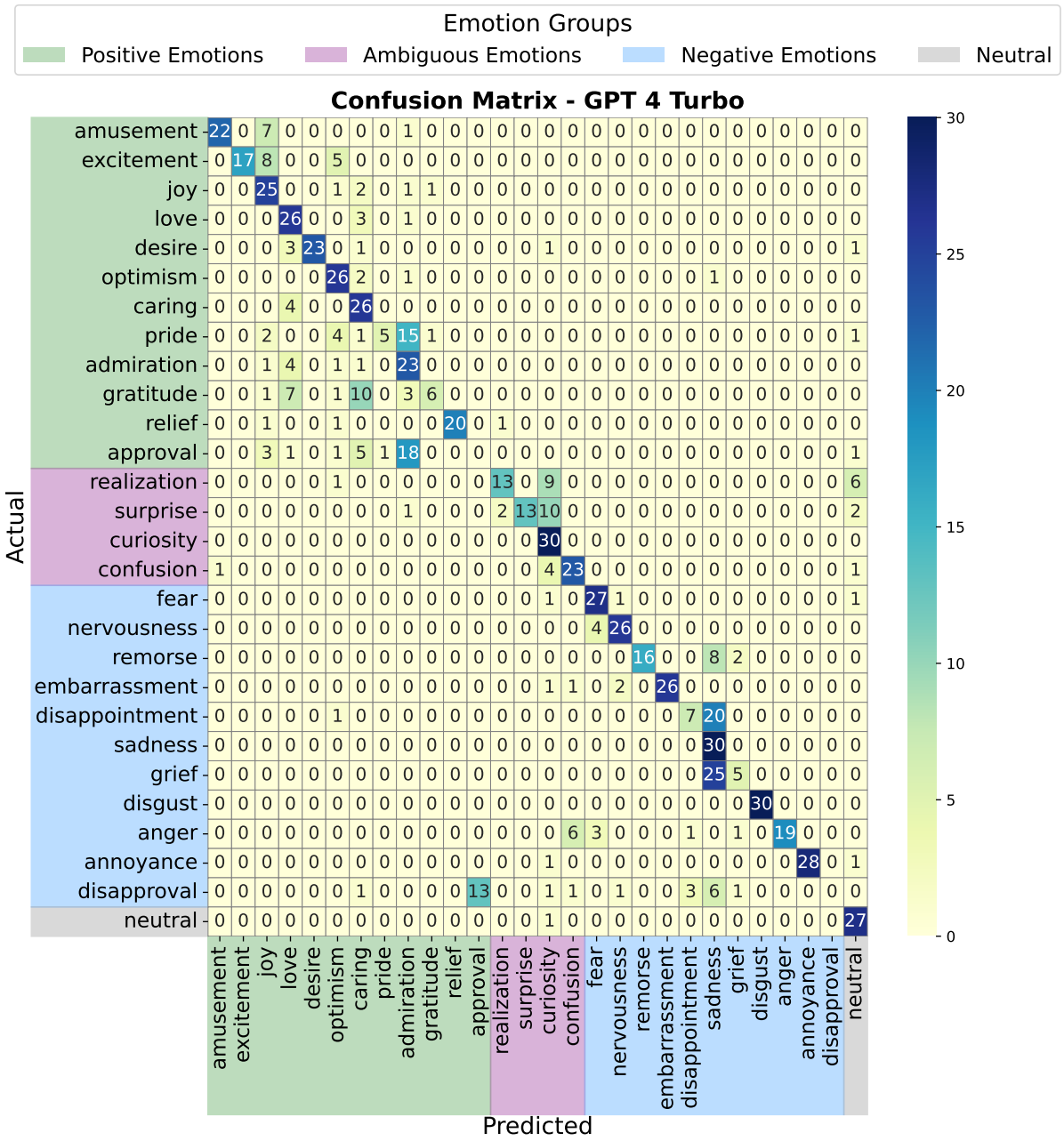


Figure 21: GPT 4: Confusion matrix of emotion signals

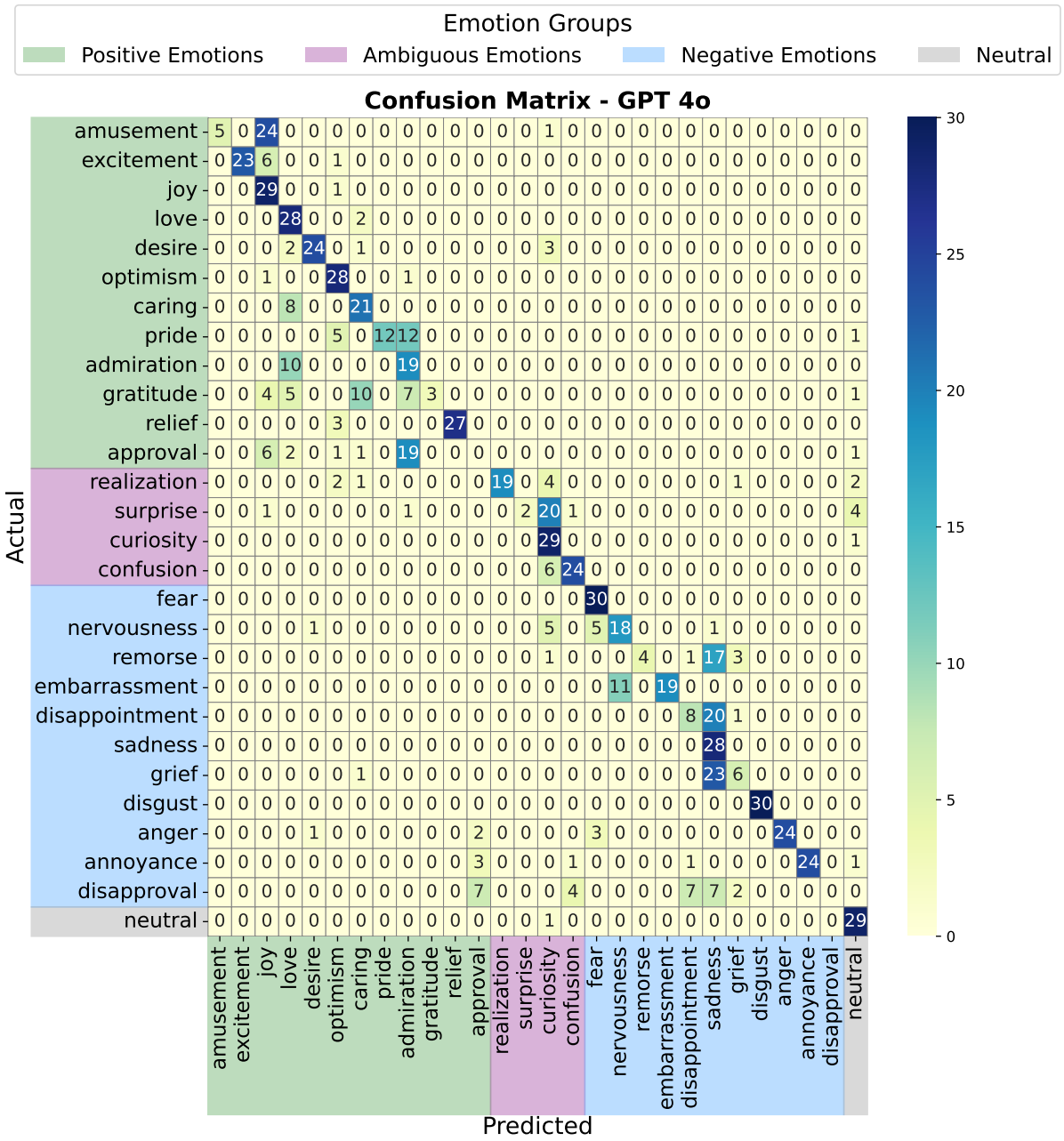


Figure 22: GPT 4o: Confusion matrix of emotion signals

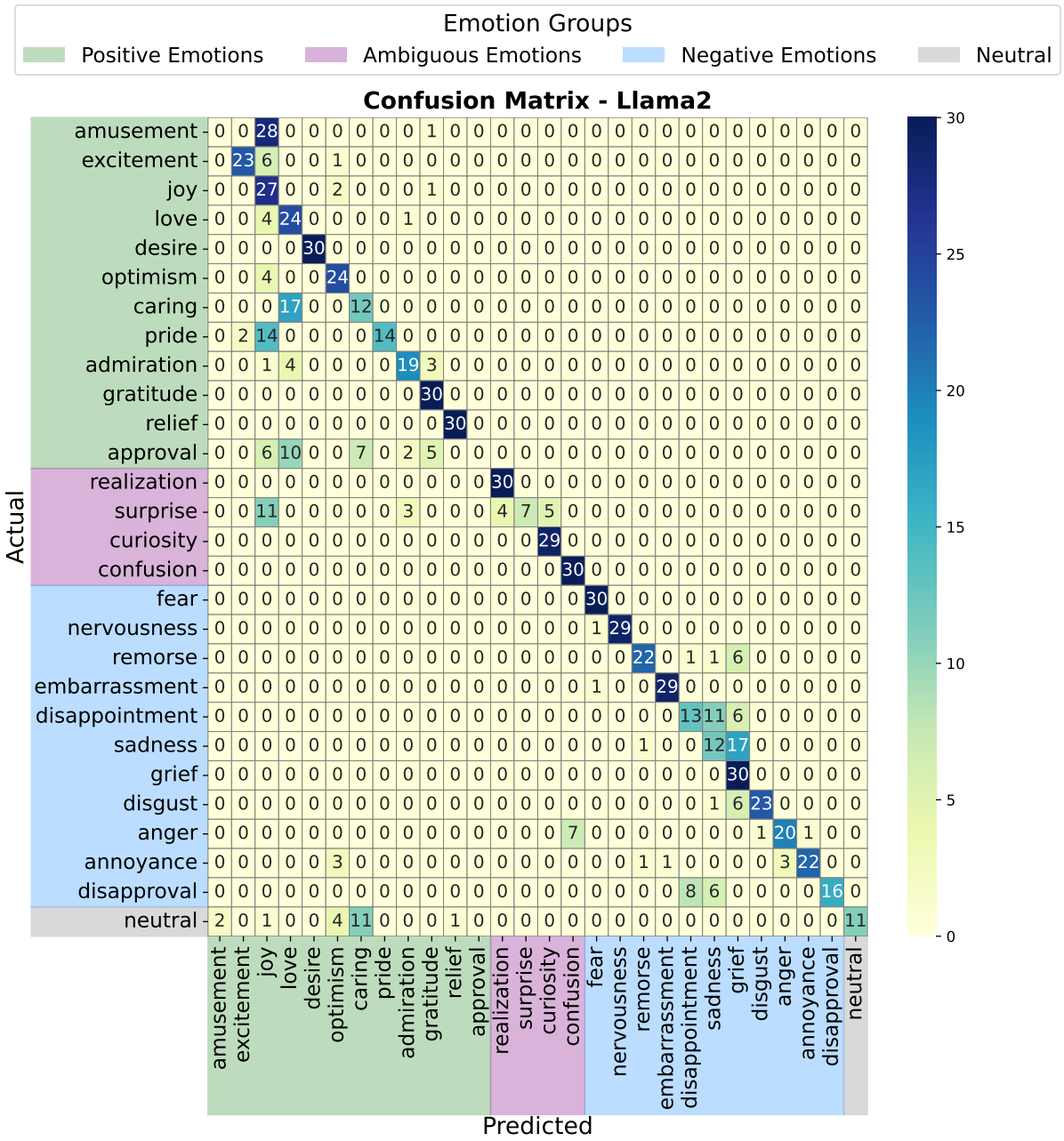


Figure 23: Llama2: Confusion matrix of emotion signals

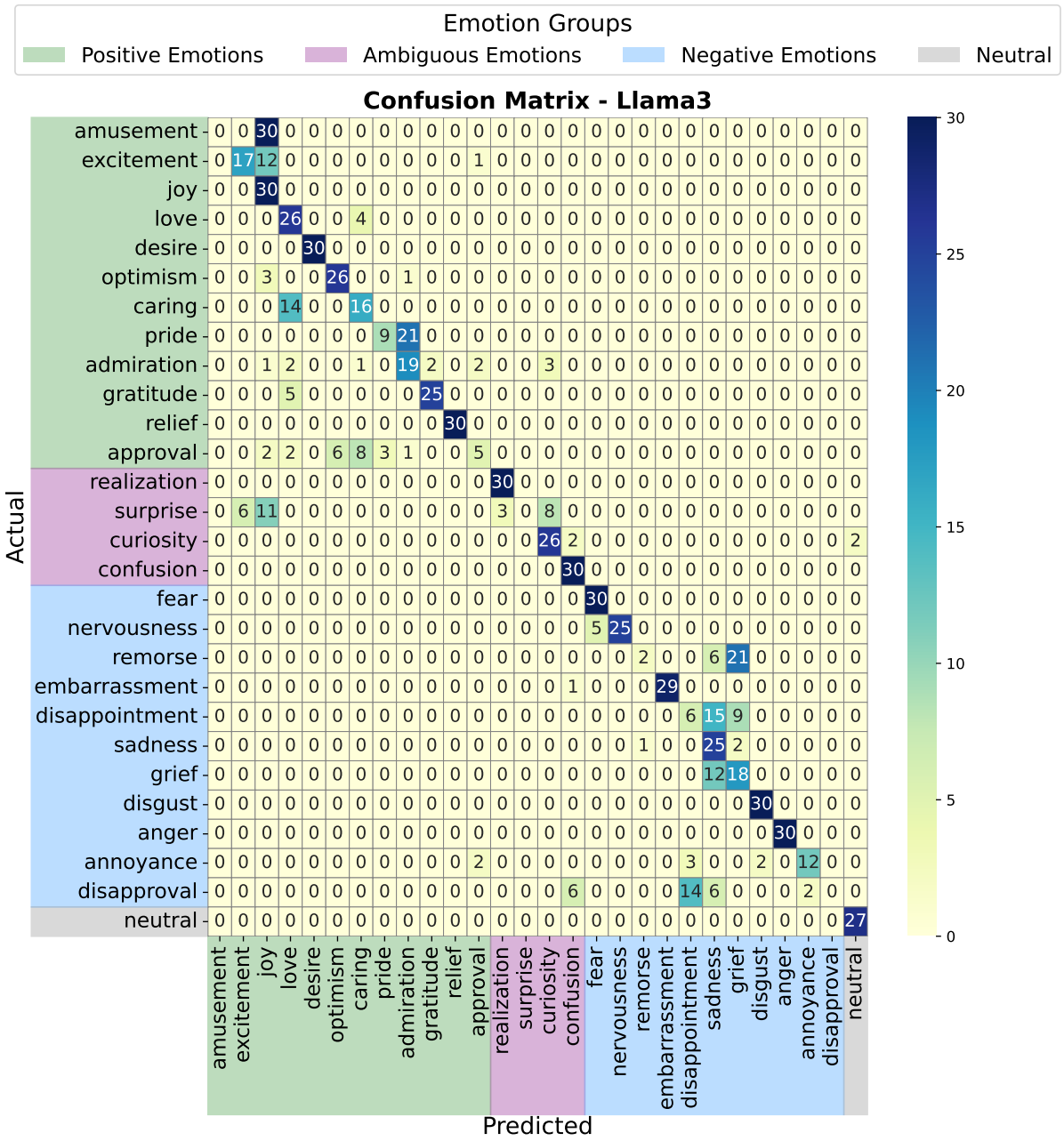
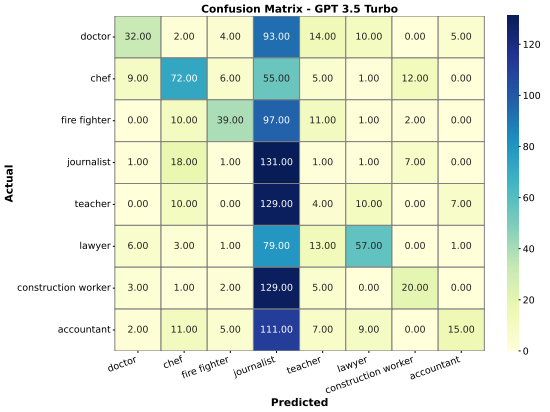
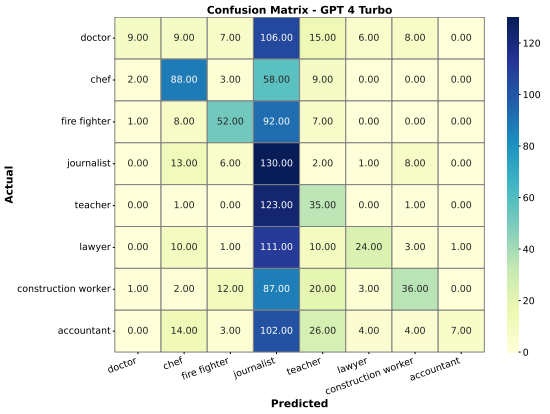


Figure 24: Llama3: Confusion matrix of emotion signals





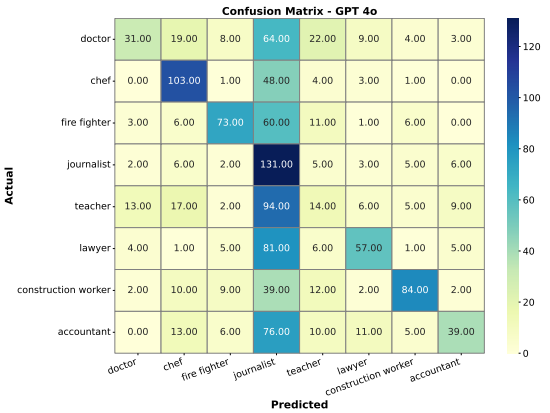
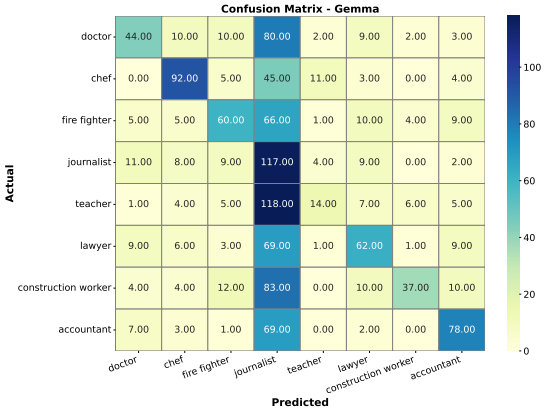
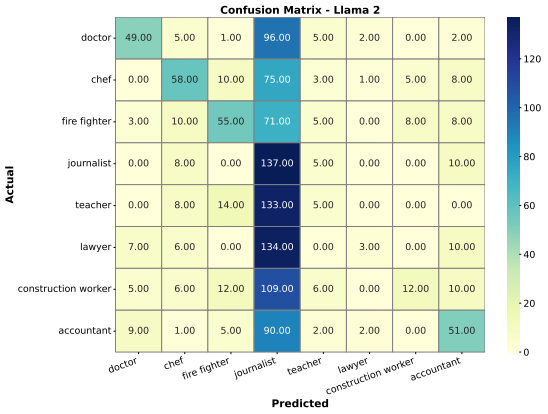


Figure 25: Confusion matrix of profession signals over a conversation across GPT models





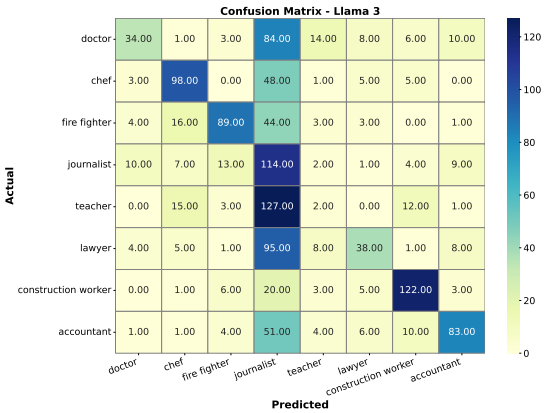


Figure 26: Confusion matrix of profession signals over a conversation across Gemma and Llama models

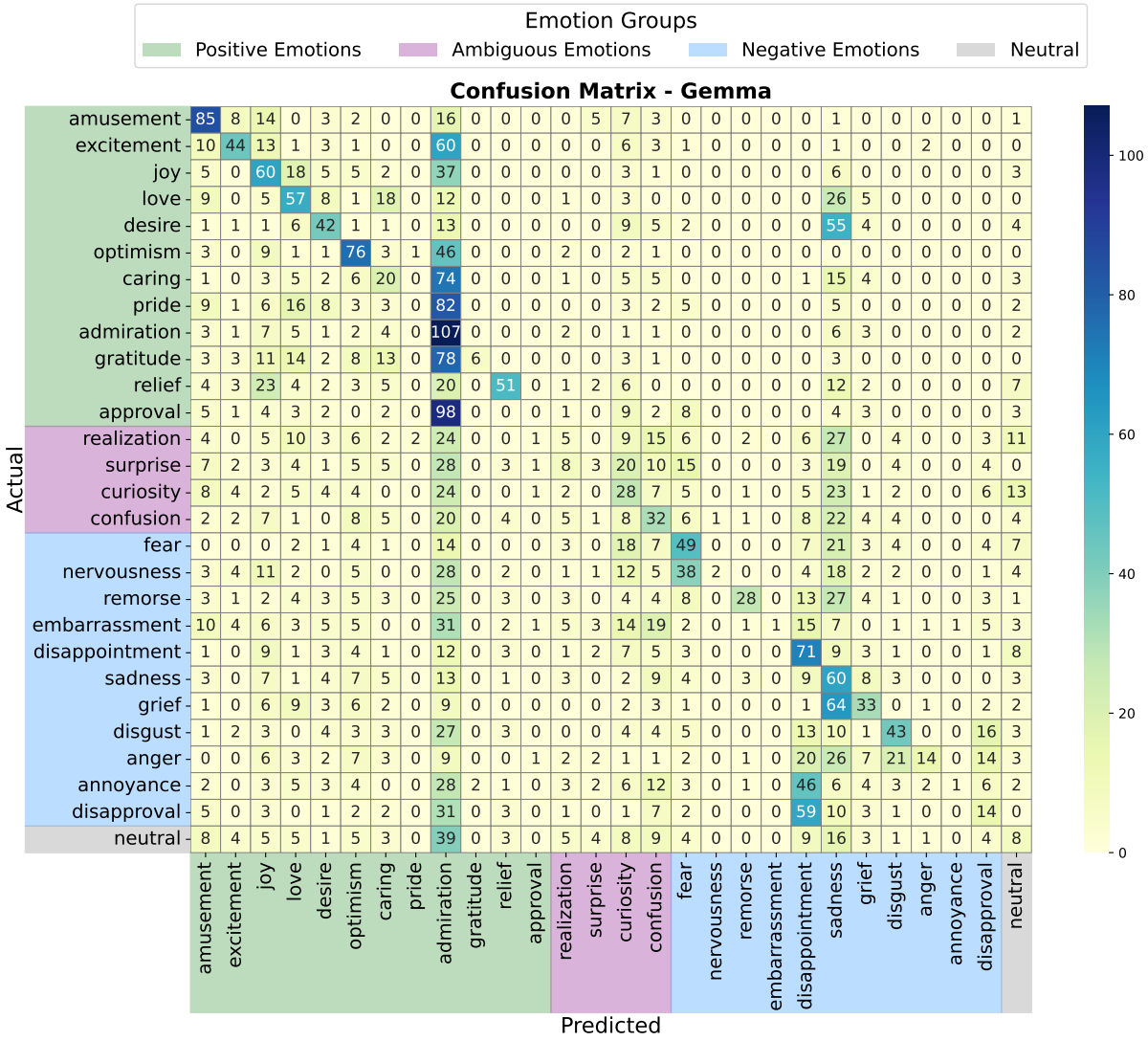


Figure 27: Gemma: Confusion matrix of emotion signals over a conversation

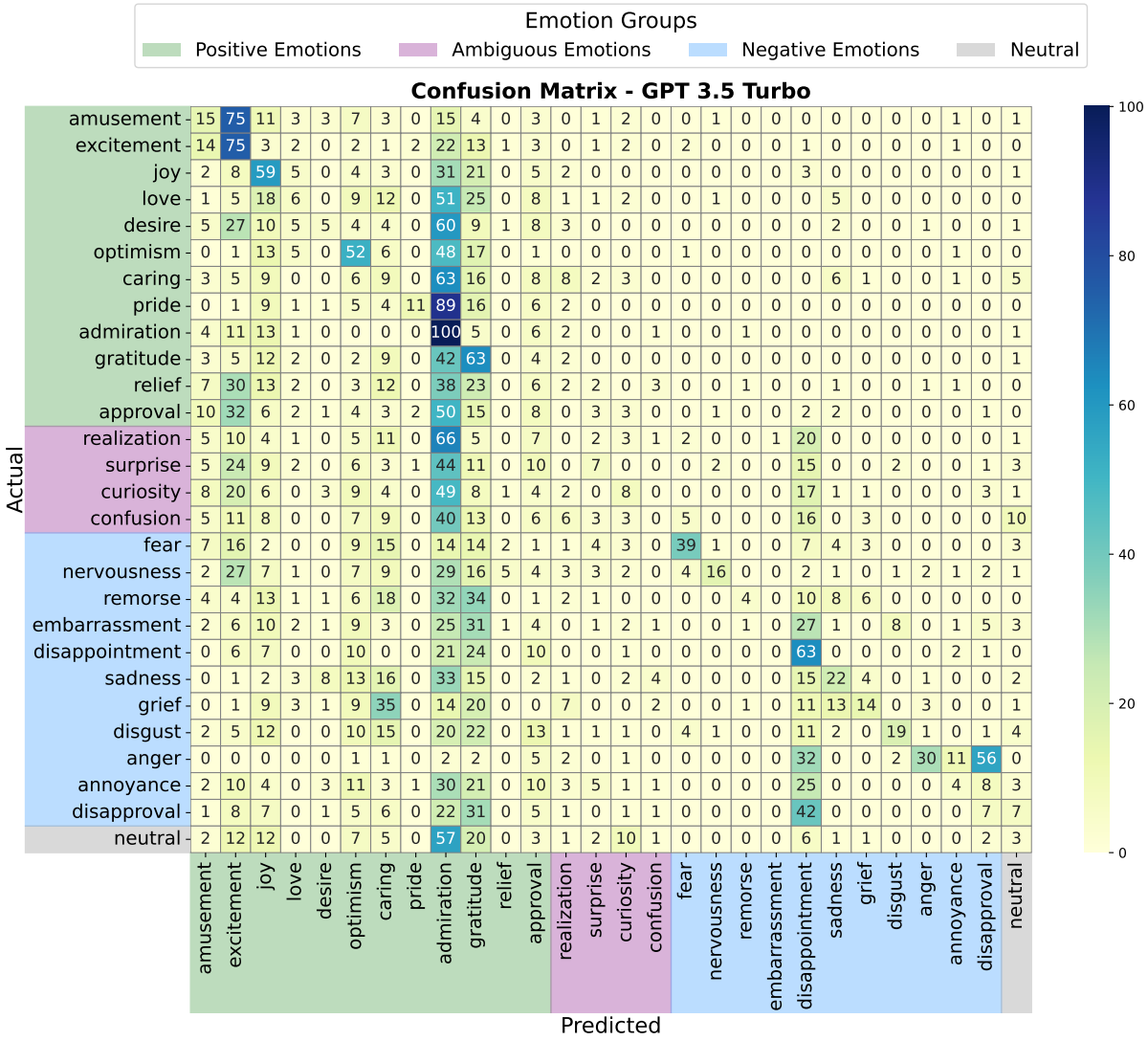


Figure 28: GPT 3.5: Confusion matrix of emotion signals over a conversation

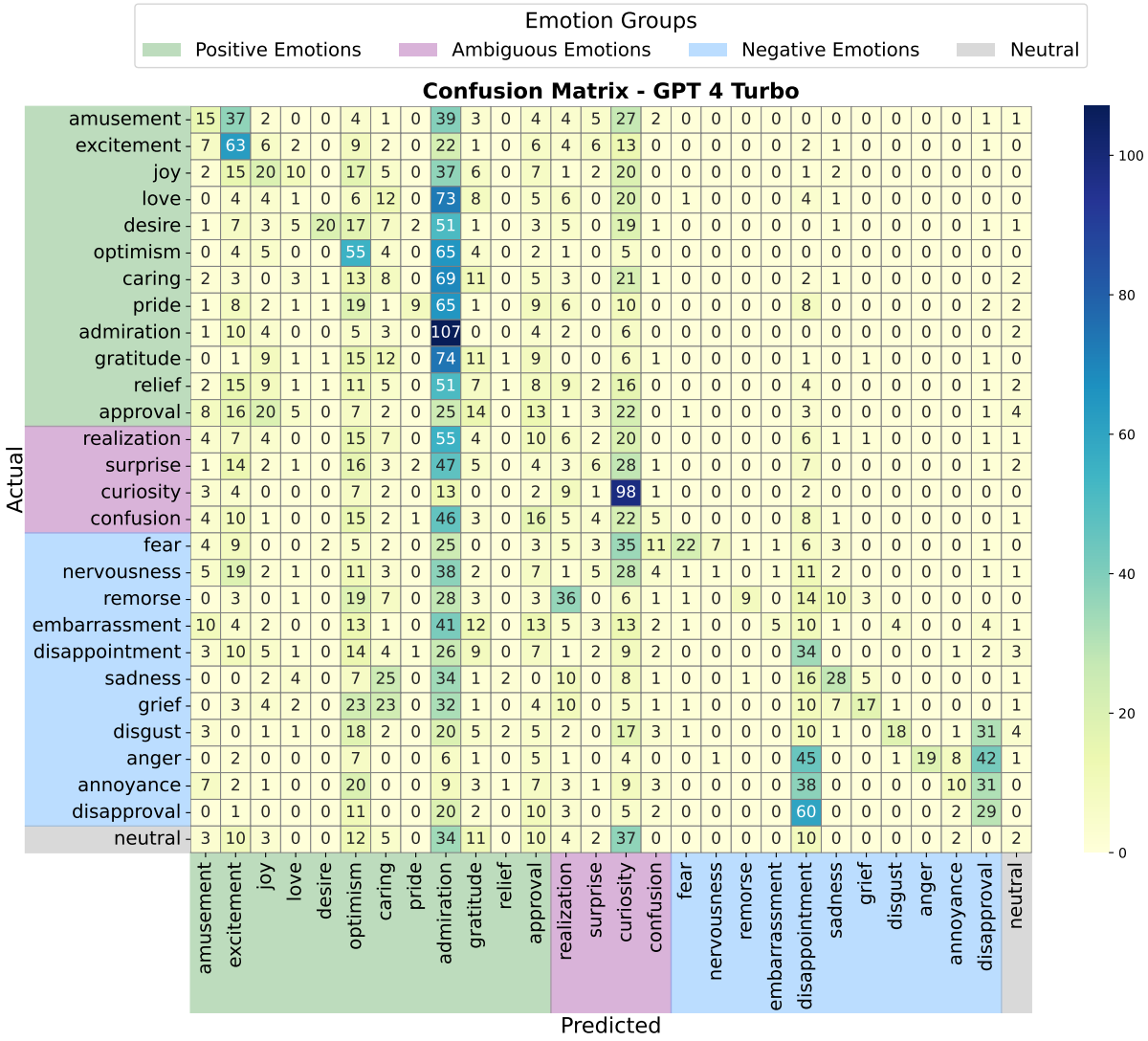


Figure 29: GPT 4: Confusion matrix of emotion signals over a conversation

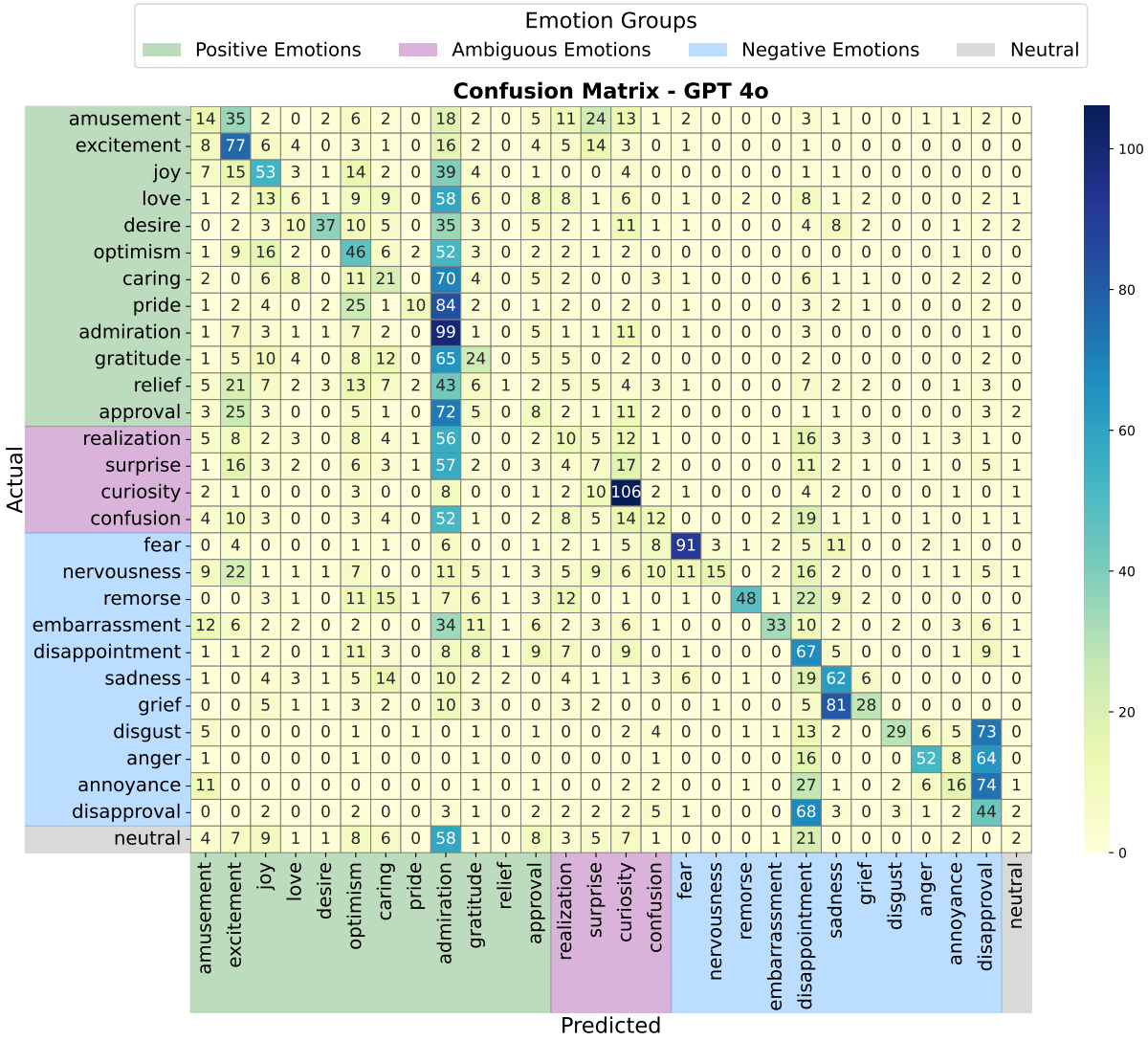


Figure 30: GPT 4o: Confusion matrix of emotion signals over a conversation

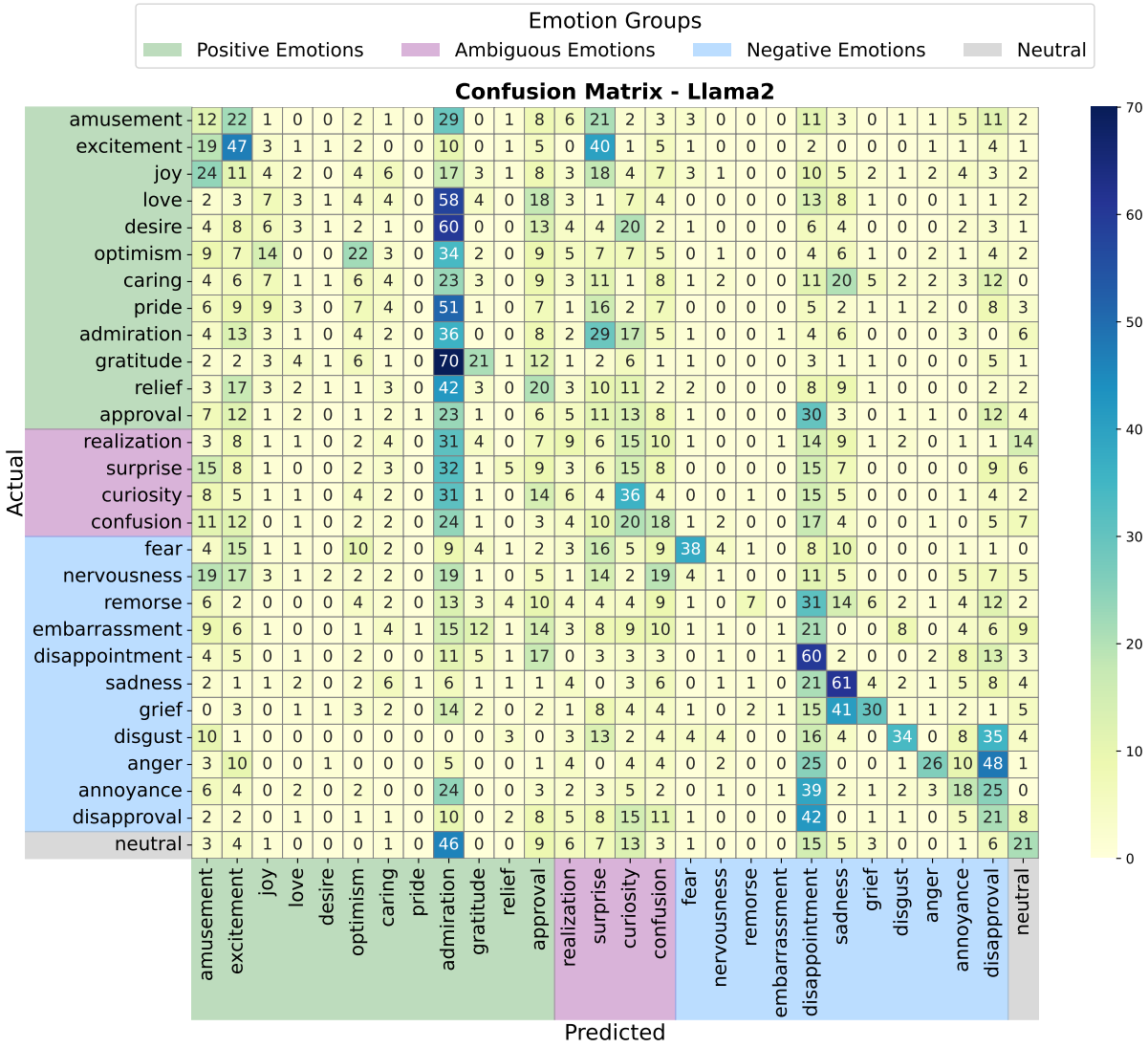


Figure 31: Llama2: Confusion matrix of emotion signals over a conversation

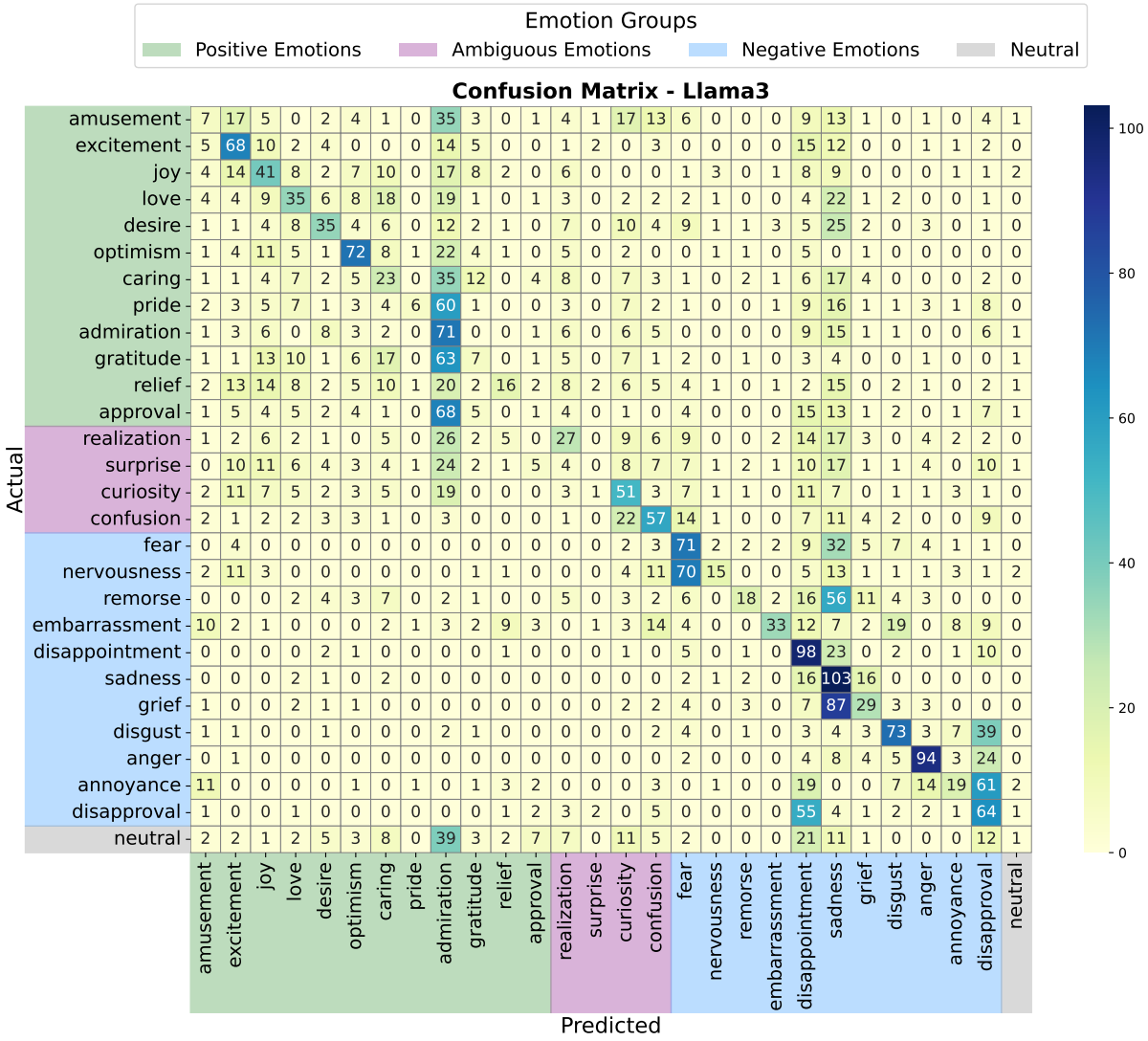


Figure 32: Llama3: Confusion matrix of emotion signals over a conversation