ExpressivityArena: Can LLMs Express Information Implicitly?

Joshua Tint[†] **Som Sagar**[†] **Aditya Taparia**[†] **Kelly Raines**[†]

Bimsara Pathiraja[†] **Caleb Liu**[†]

Ransalu Senanayake[†]
[†]Arizona State University, 1151 S Forest Ave, Tempe, AZ, USA
{jrtint, ssagar6, ataparia, kraine5, bpathir1, calebliu, ransalu}@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. 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.

1 Introduction

Large Language Models (LLMs) [1, 2] are disrupting many domains where human communication is essential, including education [1], customer support [3], legal services [4], and healthcare [5]. 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 [6, 7]. This performance is typically measured in terms of the number of errors [1], contextual understanding [7], versatility [8], problem-solving skills [5], etc. However, these tasks are often only measured as an examination of explicitly stated concepts in model output. Implicit, expressive communication largely remains unstudied. Given that much of human communication is implicit [9], expressivity may represent an important aspect of creating "human-like" output in models, improving output quality and user trust in many applications [10].

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 [11]. 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

NeurIPS 2024 Workshop on Behavioral Machine Learning.

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

indicate other information about themselves. Word choice, such as slang, may implicitly communicate one's regional background, level of education, or other identities [12].

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 [13, 14] outputs generated by various LLMs. We establish 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 analyze how expressive various LLMs are. 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.

2 Related Work

Expressivity Defined: Most methods that delve into expressivity of language models typically focus on emotions, as studied in affective computing devices [15]. This includes recognizing emotions from language or facial expressions and body language [16] or communicating emotion and personality [17, 18] in social robotics. However, this limited focus of emotions on expressivity does not capture other aspects of communication. Our study focuses on diverse aspects of expressivity, ranging from emotions to computer programming paradigms. We adapt a definition from linguistics, to term "expressivity" as the state of communicating information implicitly: *showing, not telling* [19]. 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 [20]. 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 [21].

Evaluating Large Language Models: Existing benchmarks for LLMs measure their capabilities in a variety of tasks such as mathematics [22], logical reasoning [23], and education [24]. In general, benchmarks take one of two forms: 1) automatically evaluated models by having an external LLM [25] or ensemble of LLMs [26] to act as an evaluator or 2) use human feedback to manually evaluate the model. A notable example of the latter is Chatbot Arena [27], 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 [28]. Recently, automated evaluation - or more accurately AI feedback [29, 30] - has been proposed to solve the scalability issues of Reinforcement Learning with Human Feedback (RLHF) [31]. To the best of our knowledge, previous studies on evaluation of LLMs have not focused on expressivity.

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 ⟨d⟩ which conveys ⟨s⟩. Do not explicitly mention ⟨s⟩ in your response."

We iterate this prompt for all $s \in S_{\mathcal{C}}$. The user may prompt the model: $x_{\text{in}} =$ "Write a ⟨ letter ⟩ which conveys ⟨ patriotism ⟩". 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. We define the *expressivity rate* as the rate of correct guesses in a series.

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 [32]. 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

4.1 Experiment 1: Grader Validation

Experiment 1 is designed to validate the automated grader used in ExpressivityArena, and to determine the relative accuracies of graders. LLMs have been successful in evaluating other LLMs for other tasks [31, 8, 4], 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 use GPT-4o to generate pieces of text conveying one of a set of implicit signals. The signals are chosen randomly from a list of professions (Appendix D.2.1). We used professions as our expressive signals because they don't require specialized knowledge, and they're a moderate-difficulty commonplace domain that is commonly inferred through conversation. GPT-4o 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 also use a jury grader, which aggregates responses from the LLMs: Llama3-8b, GPT-4, and Gemma [33]. We also gave these same texts to a set of human graders who were given the same task: to identify which profession was being expressed. We sought human graders through a survey distributed to 23 Arizona State University students, who were each asked to grade 5 texts. 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.

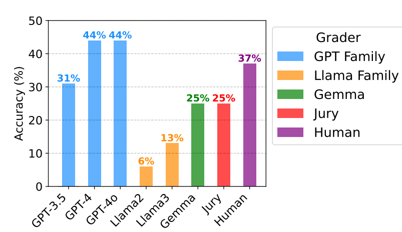


Figure 2: Accuracies of various grader types when evaluating implicitly-expressed professions.

The accuracy of each grader is shown in Fig. 2. Given that GPT-4 and GPT-4o both equally high accuracy on this experiment, we opted to use GPT-4o for our grader due to its faster performance.

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

Domain	Signal	Example Output
poem	style of Emily Dickinson	Among the clover and the nodding stems, A recluse wanders, thoughts amassed like gems...

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

Other results from this experiment were surprising; the jury [26], despite having more information from more models, was not the most accurate, possibly because two of its constituent models substantially underperformed its other constituent model, GPT-4o. That GPT-4 and GPT-4o outperforms even a set of human graders is also unexpected, and could be due to the greater breadth of dialogue that the models have access to. Human graders tended to be consistent, with an average Fleiss' Kappa of 0.83.

4.2 Experiment 2: Single-Prompt Scenarios

The purpose of experiment 2 is to answer the questions: Are LLMs capable of exhibiting expressivity? we consider single-prompt scenarios—the user prompts only once and $f_{\text{test}}(\cdot)$ generates a single response. We evaluate two domains, these being poetry generation and code generation.

4.2.1 Poetry Generation

Poetry is typically thought of as a highly expressive domain. Studying poetry for this analysis allows us to better grasp whether LLMs have the capacity to match the expressivity of the most expressive humans. For the poetry domain, we study two different kinds of signals, emotion and writing style. For the emotion category, we use the set of emotions from the GoEmotions dataset [34] as our set of signals. 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.

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

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 D.2.1. 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 overall in expressing poets' styles than emotions. The worst performance was from Gemma with an expressivity rate of 0.53, and the best was from GPT-4 with an expressivity rate of 0.70. For several models, when asked to give a poem in the style of a female poet such as Elizabeth Barrett Browning 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. Complete confusion matrices for both tests can be found in the Appendix D.

4.2.2 Code Generation

As opposed to poetry, programming is not traditionally considered as an expressive domain. However, in contrast to poems, programs can be formally checked for correctness against a specification, allowing us to better understand if introducing expressive constraints interferes with correctness or functionality in model output. We studied expressivity in two subcategories for program generation: skill level and programming style. 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 result 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 this experiment, the expressive signals were "functional," "procedural," "object-oriented," and "array-oriented," four 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 average expressivity rate 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 of how frequently each label was assigned to each model's output given its assigned prompt in each experiment is available in the Appendix D. 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 randomly. Therefore, in answering the question "Are LLMs capable of exhibiting expressivity?," we must conclude that LLMs struggle at expressivity in the context of code compared to highly expressive domains like poetry.

5 Conclusion

ExpressivityArena served as a platform to analyze expressivity of LLMs in single-prompt response generations. Based on the results from our experiments, LLMs have shown to be capable of some level of expressivity. However, there is still much room for improvement since the accuracy of LLMs trends around 30-60%. This performance may reflect underlying biases related to race, age, gender, or other factors that are underrepresented in models which become more apparent in creative and logical fields tasks. 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.

References

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

[2] 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. Llama: Open and efficient foundation language models, 2023. cite arxiv:2302.13971.

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

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

[5] 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. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

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

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

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

[9] Ross A. Knepper, Christoforos I. Mavrogiannis, Julia Proft, and Claire Liang. 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, 2017. Association for Computing Machinery.

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

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

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

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

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

[15] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[16] Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia, May 2024. ELRA and ICCL.

[17] Zining Wang, Paul Reisert, Eric Nichols, and Randy Gomez. 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, 2024. Association for Computing Machinery.

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

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

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

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

[22] 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. Evaluating language models for mathematics through interactions. *arXiv preprint arXiv:2306.01694*, 2023.

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

[24] Wei Dai, Jionghao Lin, Hua Jin, Tongguang Li, Yi-Shan Tsai, Dragan Gašević, and Guanliang Chen. 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, 2023.

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

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

[27] 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. Chatbot arena: An open platform for evaluating llms by human preference, 2024.

[28] 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. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3), mar 2024.

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

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

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

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

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

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

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

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

Appendix

A 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.

B 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.

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.

In this study, we use the multiple choice metric, which has been shown to be nonlinear and discontinuous for NLP tasks [35]. 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.

Autonomated evaluation is used within the study. However, biases or unexpected behavior within the judging model(s) could lead to incorrect evaluations that differ from human judgment [25]. Thus, we conducted experiments with human evaluators to confirm that automated evaluation is viable for this task.

C Discussion

Based on the results of experiment 1, we uncovered that the best-performing automated graders were about equally proficient in identifying implicit signals as humans. Automated graders even surpassed the performance of humans at some points. This may be due to various reasons: automated graders have a better understanding of associations that LLMs make, or perhaps human graders experience mental fatigue while evaluating the dialogue [36]. The highest performing models were GPT-4 and GPT-4o, which may be due to their larger size; GPT-4 has over a trillion parameters, compared to many of the other models, which had under 10 billion [1]. The high overall performance of GPT-4 and GPT-4o led us to conclude that automated graders would be appropriate for the following experiments in which we use them to measure the expressiveness of LLMs.

In experiment 2, we utilized automated graders to evaluate the expressiveness of LLMs. Broken up into two subsections — logical and creative generative outputs — we chose coding and poems, respectively, to represent these areas. In more logical areas such as code, expressiveness 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 the poetry domain, there were significant levels of confusion between female poets which impacted the accuracy of each model. This suggests that bias may negatively impact expressivity. Poets such as Emily Dickinson and Sappho come from greatly differing backgrounds which influenced their work and, thus style of writing. However, as the model confuses the two female poets, it seems that the model has overgeneralized the main themes explored in their writing based on their sex, and does not exhibit enough expressiveness to differentiate them. Models' training data may have underrepresented female poets, leading to this generalization. In addition to writing styles, certain emotions in experiment 2 were frequently confused, typically ones with similar semantics (e.g., confusing one positive emotion for another). 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.

D Additional Results

D.1 Experiment 1: LLM as a grader

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. 3).

D.2 Experiment 2: Single-prompt scenarios

Code skill Signals: Fig. 4 shows the confusion matrix of provided and predicted code skill signals for different LLM models.

Paradigms Signals: Fig. 5 shows the confusion matrix of provided and predicted paradigms signals for different LLM models.

Poets' Signals: Fig. 6, 7, 8, 9, 10, and 11 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. 12, 13, 14, 15, 16, and 17 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.

D.2.1 Signal Categories

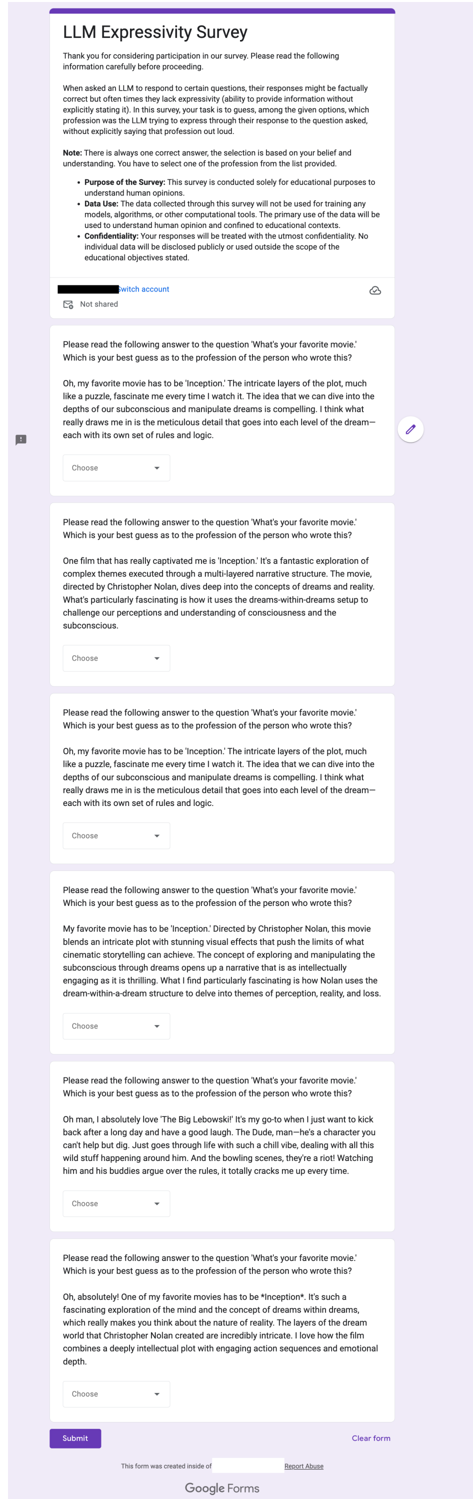
Emotions: The list of emotions used is as follows:

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

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



LLM Expressivity Survey

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.

[redacted]witch account

Not shared

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

Oh, my favorite movie has to be 'Inception.' The intricate layers of the plot, much like a puzzle, fascinate me every time I watch it. The idea that we can dive into the depths of our subconscious and manipulate dreams is compelling. I think what really draws me in is the meticulous detail that goes into each level of the dream—each with its own set of rules and logic.

Choose

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

One film that has really captivated me is 'Inception.' It's a fantastic exploration of complex themes executed through a multi-layered narrative structure. The movie, directed by Christopher Nolan, dives deep into the concepts of dreams and reality. What's particularly fascinating is how it uses the dreams-within-dreams setup to challenge our perceptions and understanding of consciousness and the subconscious.

Choose

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

Oh, my favorite movie has to be 'Inception.' The intricate layers of the plot, much like a puzzle, fascinate me every time I watch it. The idea that we can dive into the depths of our subconscious and manipulate dreams is compelling. I think what really draws me in is the meticulous detail that goes into each level of the dream—each with its own set of rules and logic.

Choose

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

My favorite movie has to be 'Inception.' Directed by Christopher Nolan, this movie blends an intricate plot with stunning visual effects that push the limits of what cinematic storytelling can achieve. The concept of exploring and manipulating the subconscious through dreams opens up a narrative that is as intellectually engaging as it is thrilling. What I find particularly fascinating is how Nolan uses the dream-within-a-dream structure to delve into themes of perception, reality, and loss.

Choose

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

Oh man, I absolutely love 'The Big Lebowski!' It's my go-to when I just want to kick back after a long day and have a good laugh. The Dude, man—he's a character you can't help but dig. Just goes through life with such a chill vibe, dealing with all this wild stuff happening around him. And the bowling scenes, they're a riot! Watching him and his buddies argue over the rules, it totally cracks me up every time.

Choose

Please read the following answer to the question 'What's your favorite movie.' Which is your best guess as to the profession of the person who wrote this?

Oh, absolutely! One of my favorite movies has to be *Inception*. It's such a fascinating exploration of the mind and the concept of dreams within dreams, which really makes you think about the nature of reality. The layers of the dream world that Christopher Nolan created are incredibly intricate. I love how the film combines a deeply intellectual plot with engaging action sequences and emotional depth.

Choose

Submit

Clear form

This form was created inside of Report Abuse

Google Forms

Figure 3: An unfilled example survey.

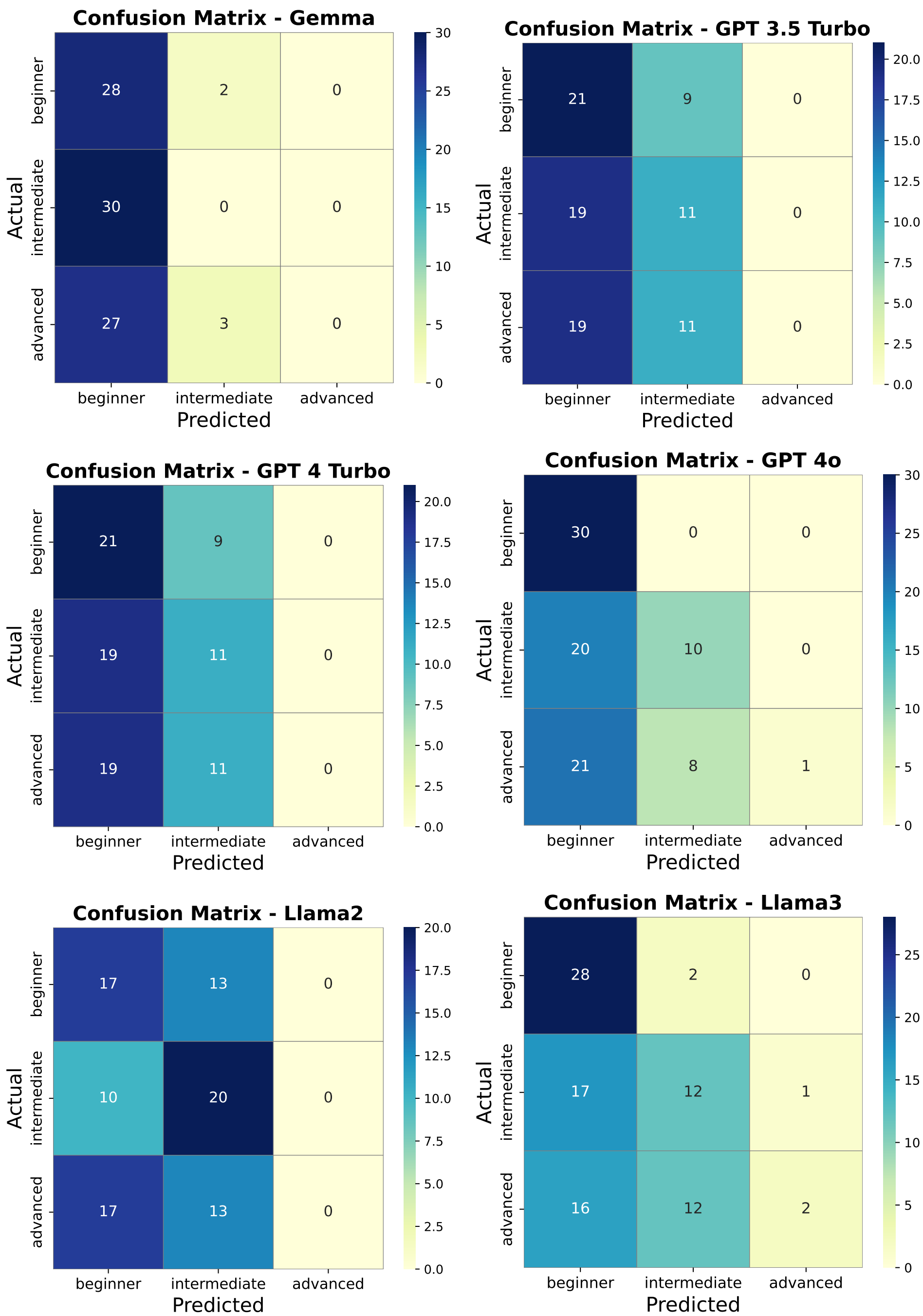


Figure 4: Confusion matrix of provided code skill signals and predicted code skill signals

33. Li Bai
34. Jalal Al-Din Rumi

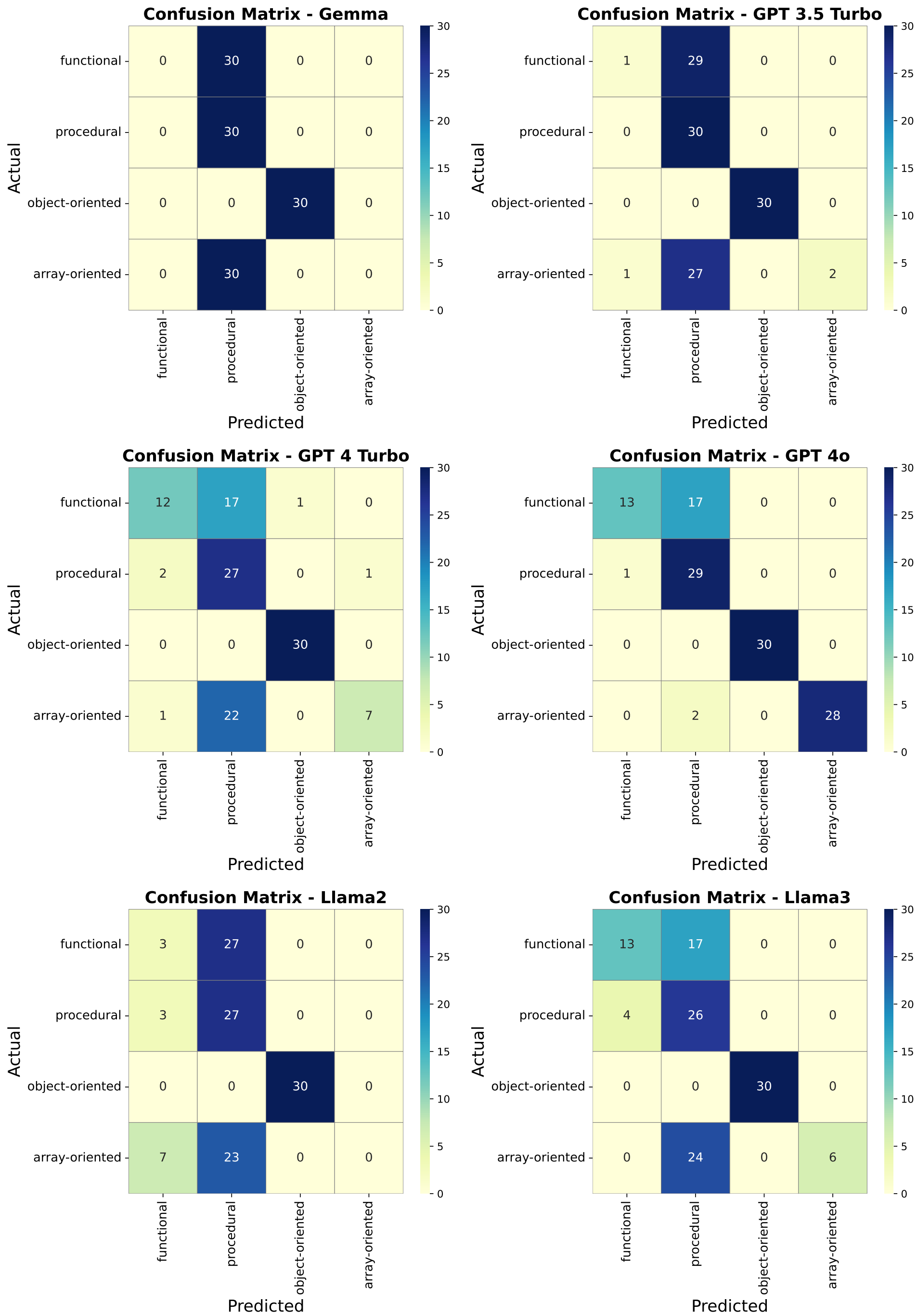


Figure 5: Confusion matrix of provided paradigms signals and predicted paradigms signals

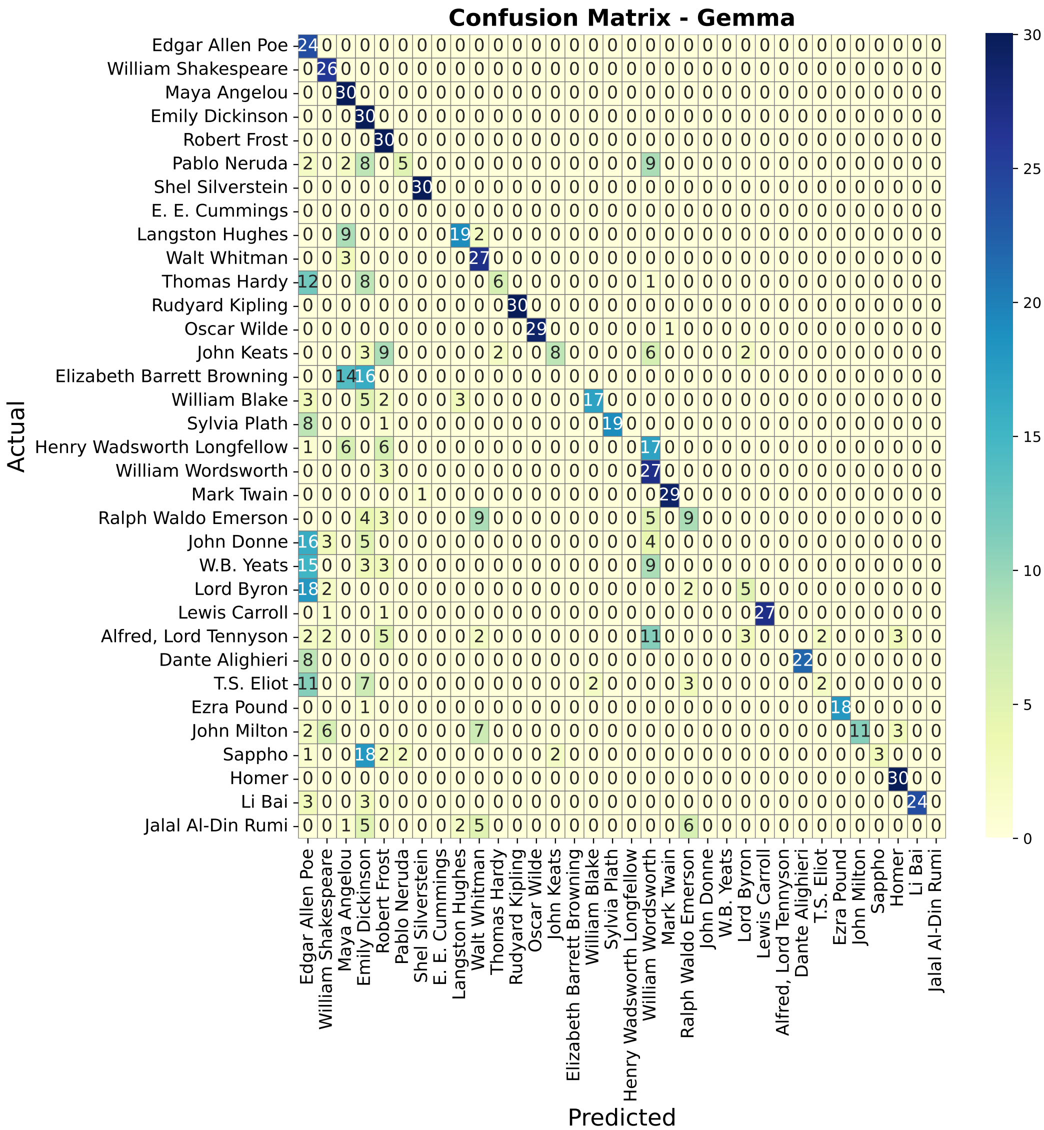


Figure 6: Gemma: Confusion matrix of poets' signals

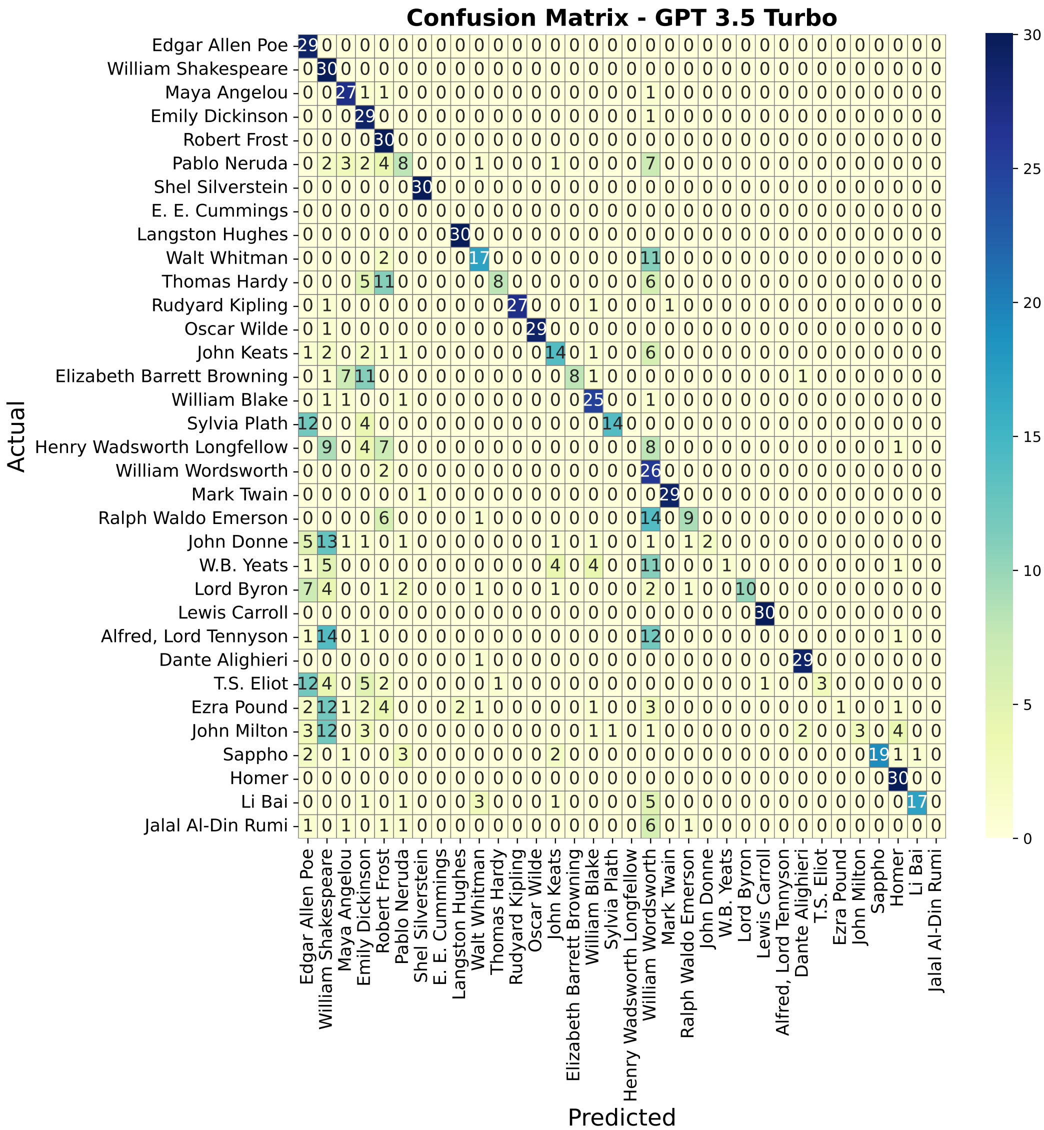


Figure 7: GPT 3.5: Confusion matrix of poets' signals

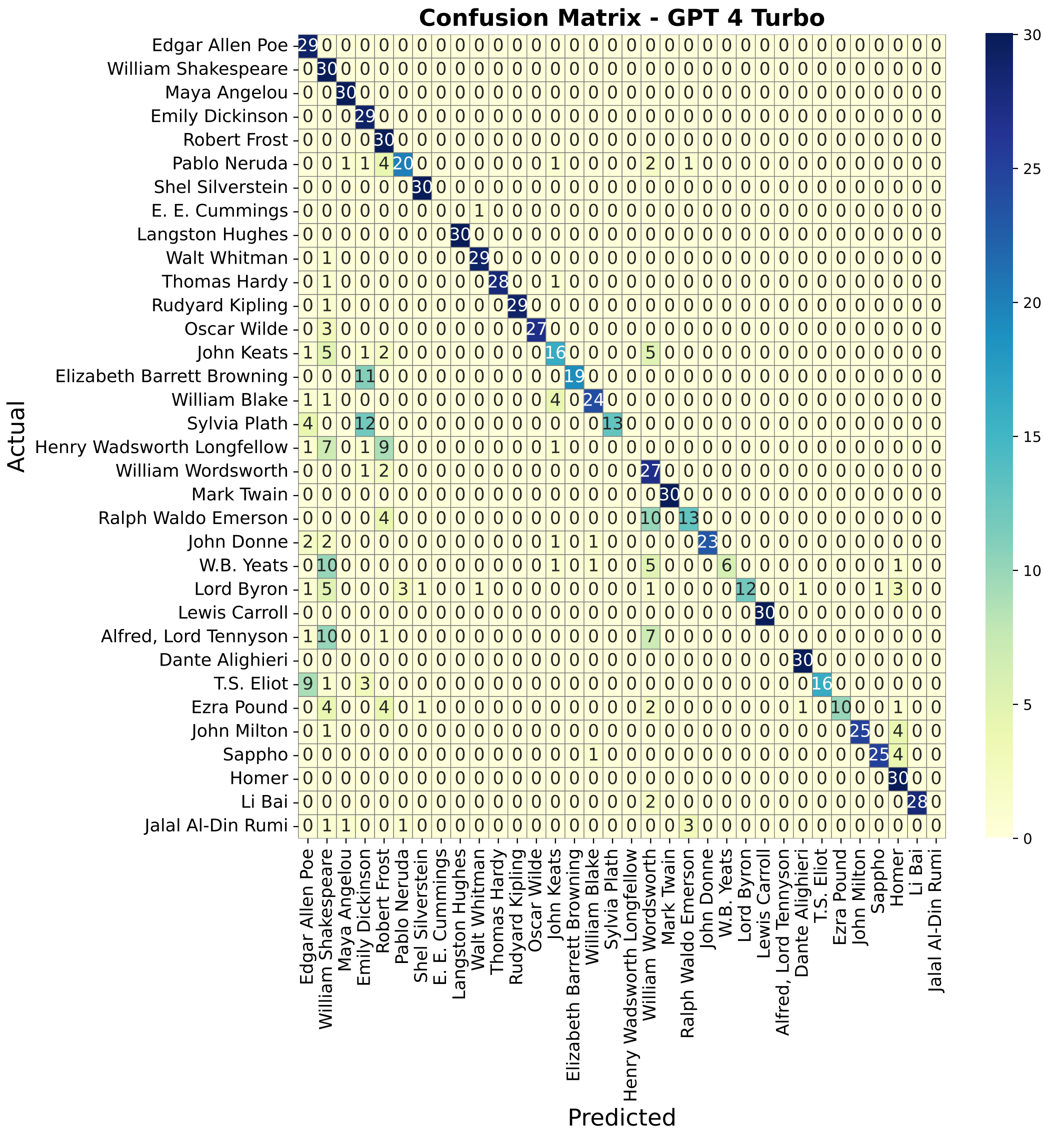


Figure 8: GPT 4: Confusion matrix of poets' signals

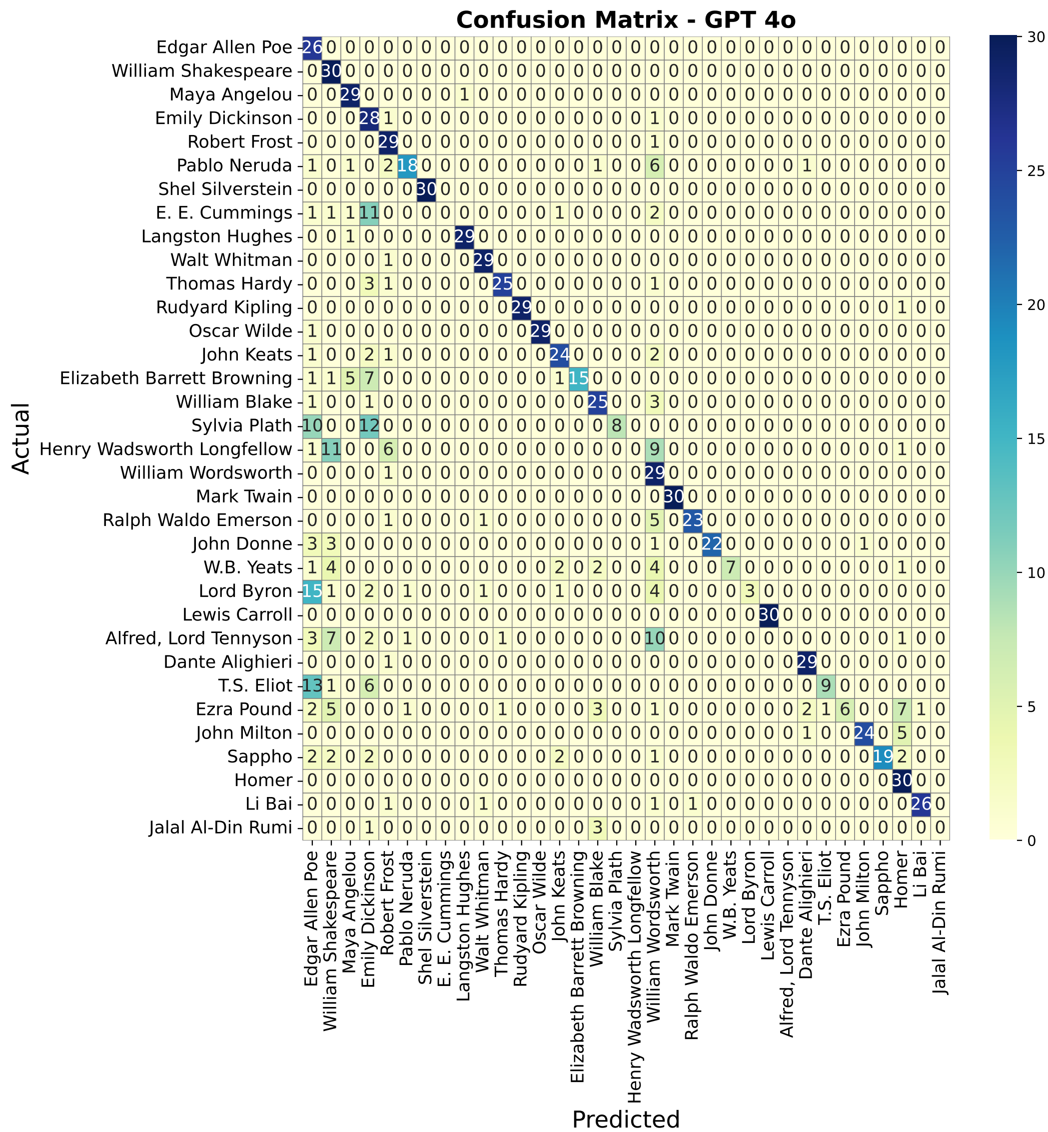


Figure 9: GPT 4o: Confusion matrix of poets' signals

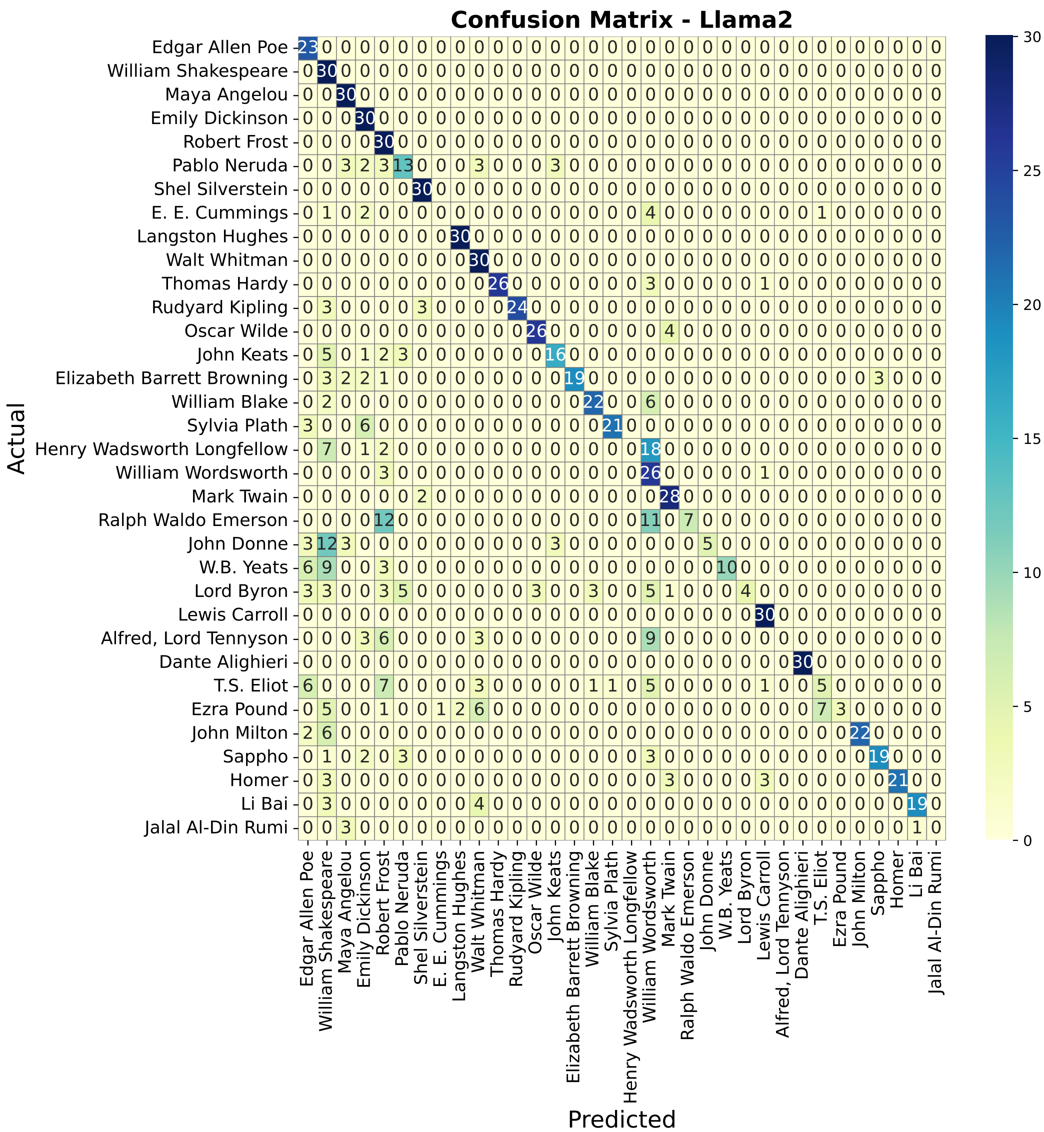


Figure 10: Llama2: Confusion matrix of poets' signals

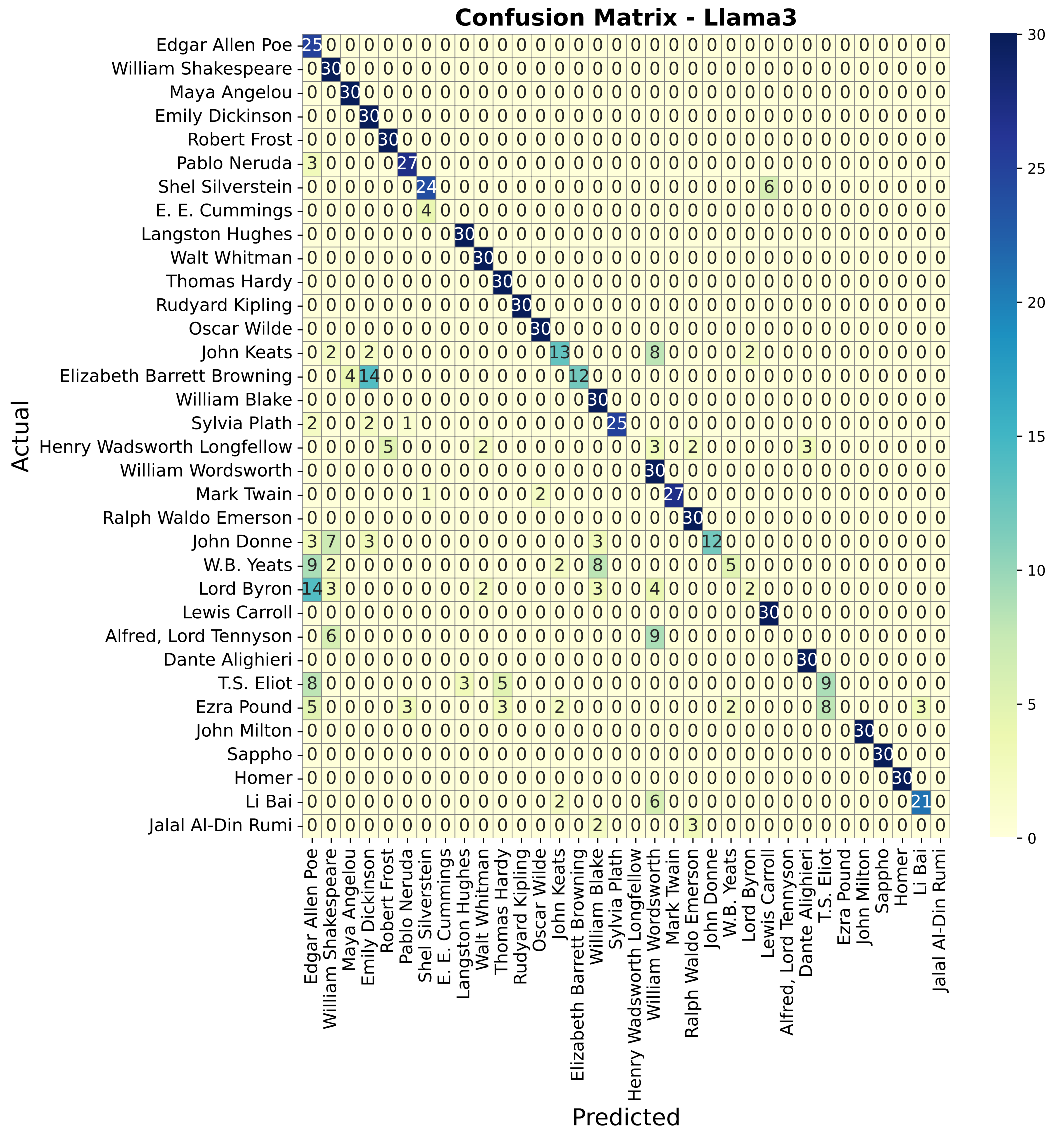


Figure 11: Llama3: Confusion matrix of poets' signals

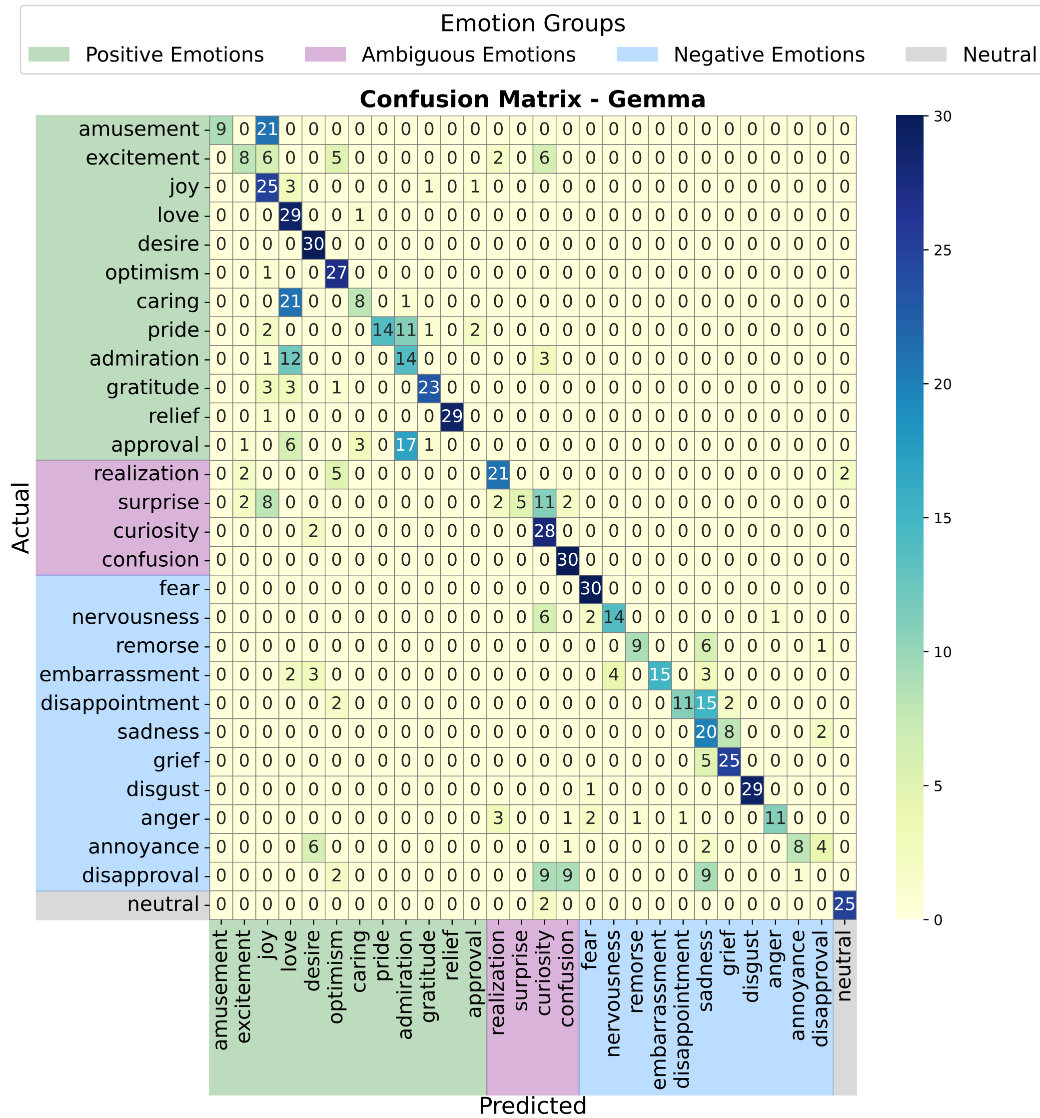


Figure 12: Gemma: Confusion matrix of emotion signals

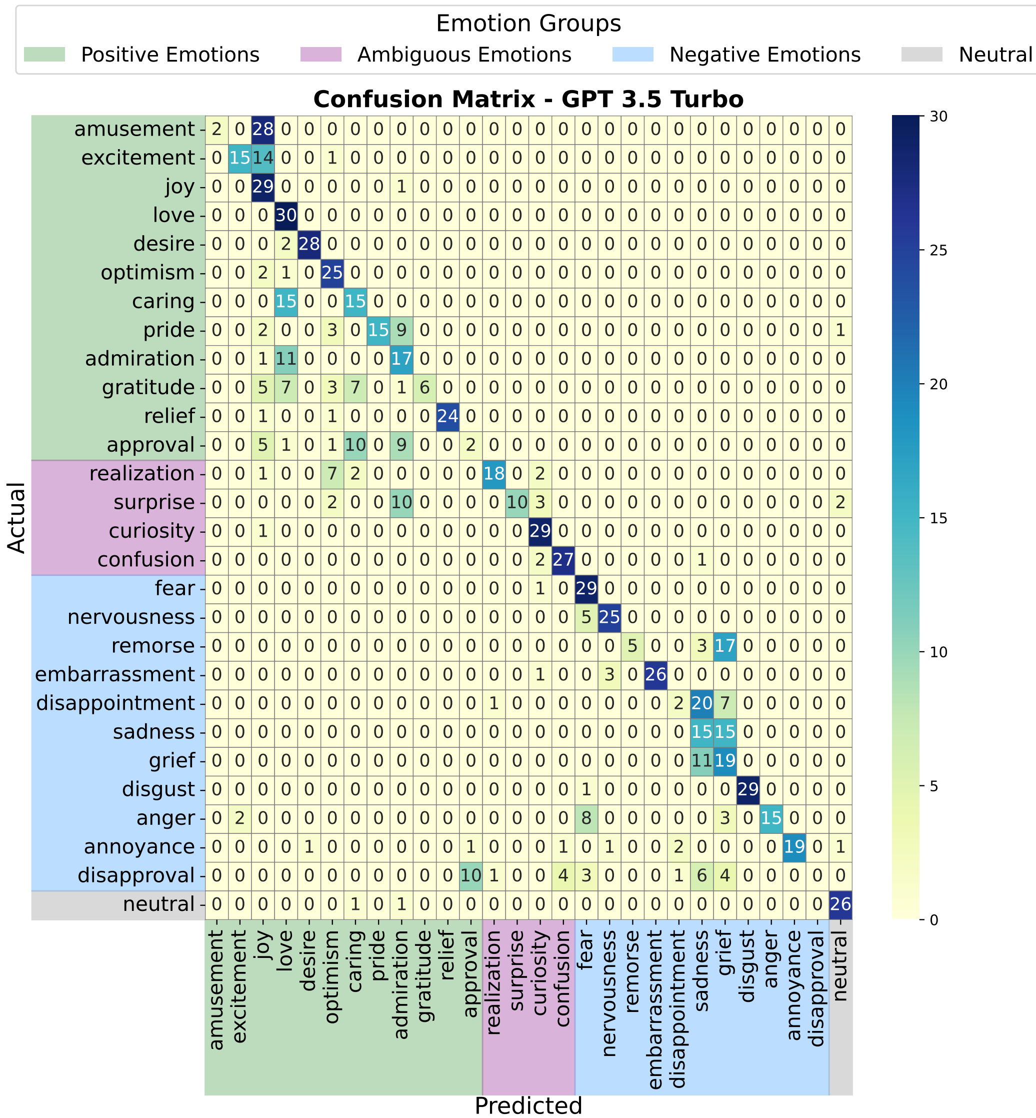


Figure 13: GPT 3.5: Confusion matrix of emotion signals

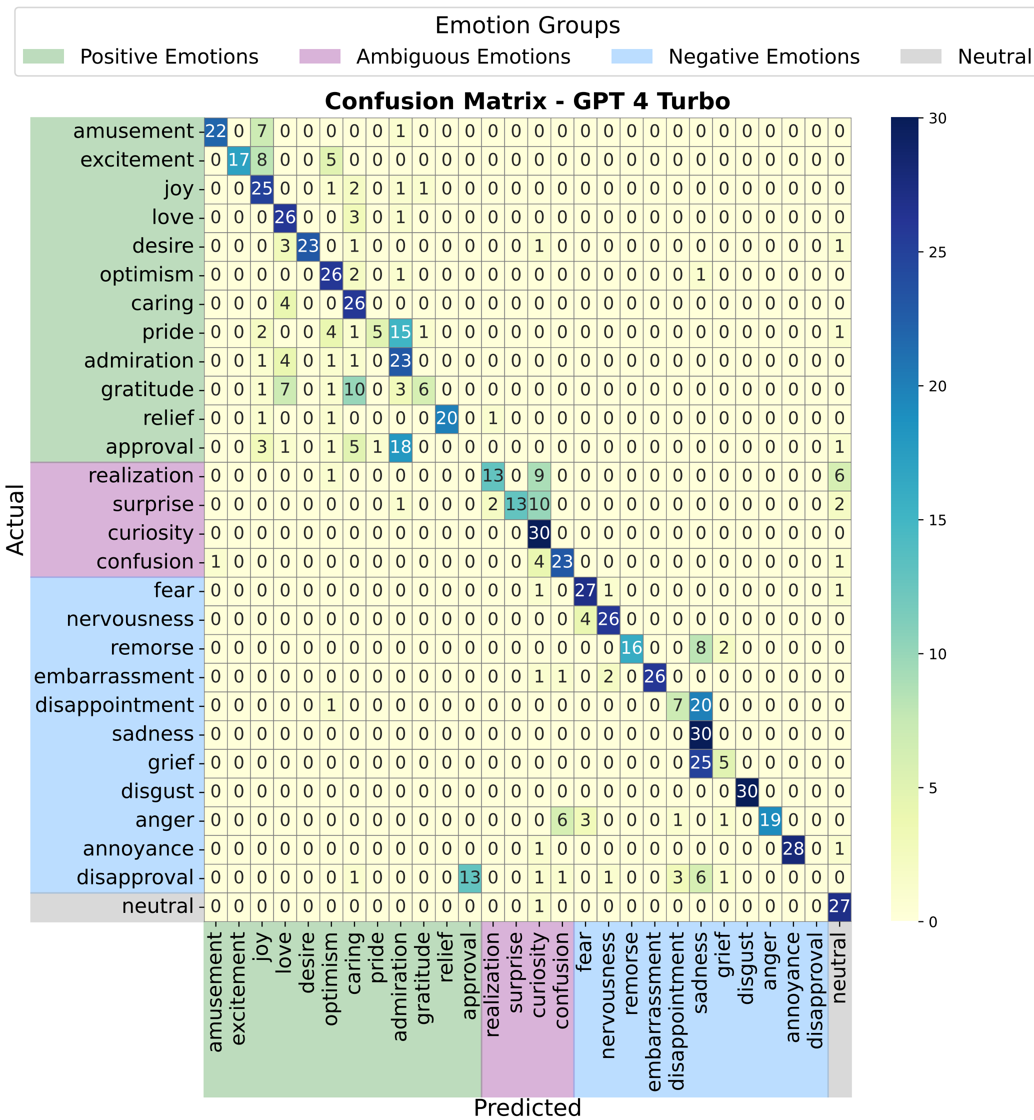


Figure 14: GPT 4: Confusion matrix of emotion signals

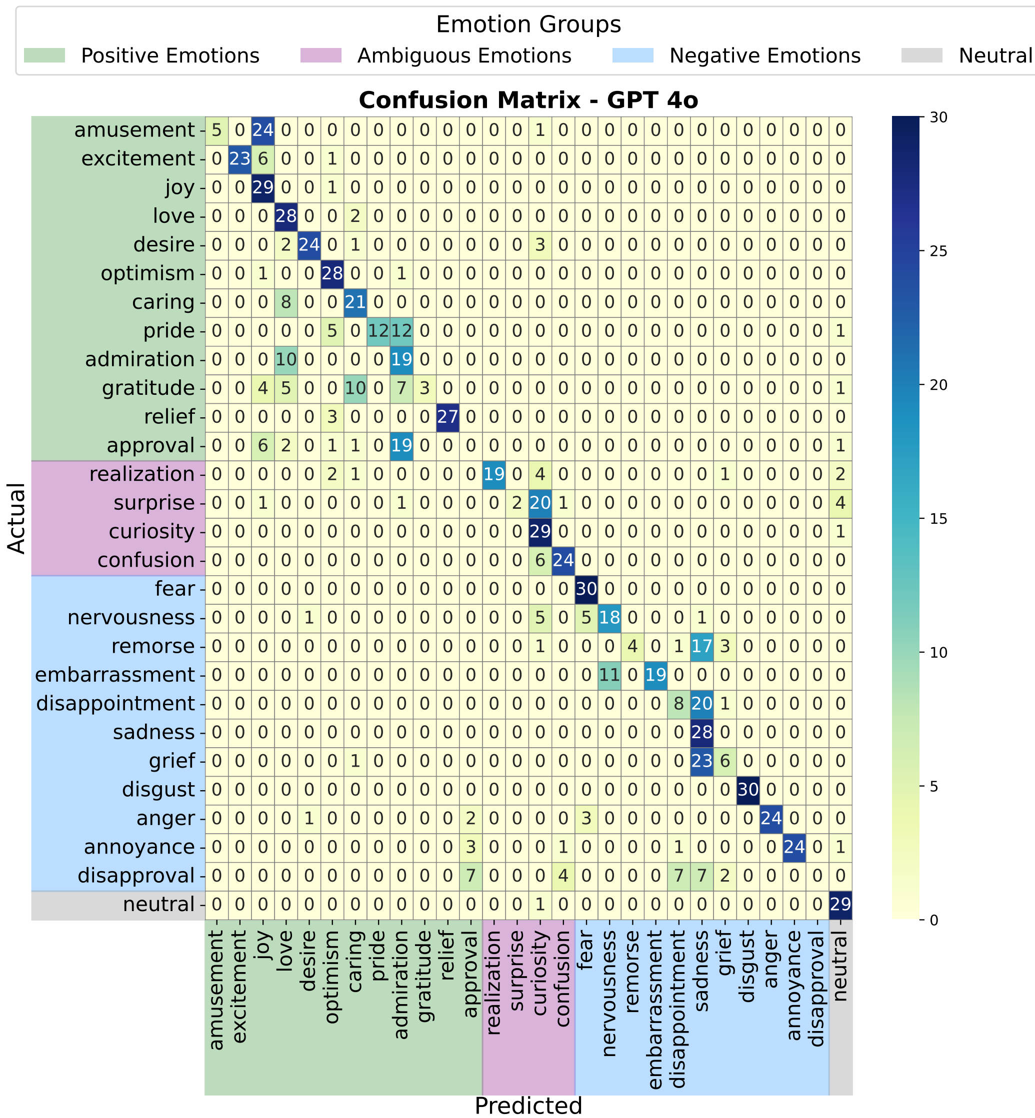


Figure 15: GPT 4o: Confusion matrix of emotion signals

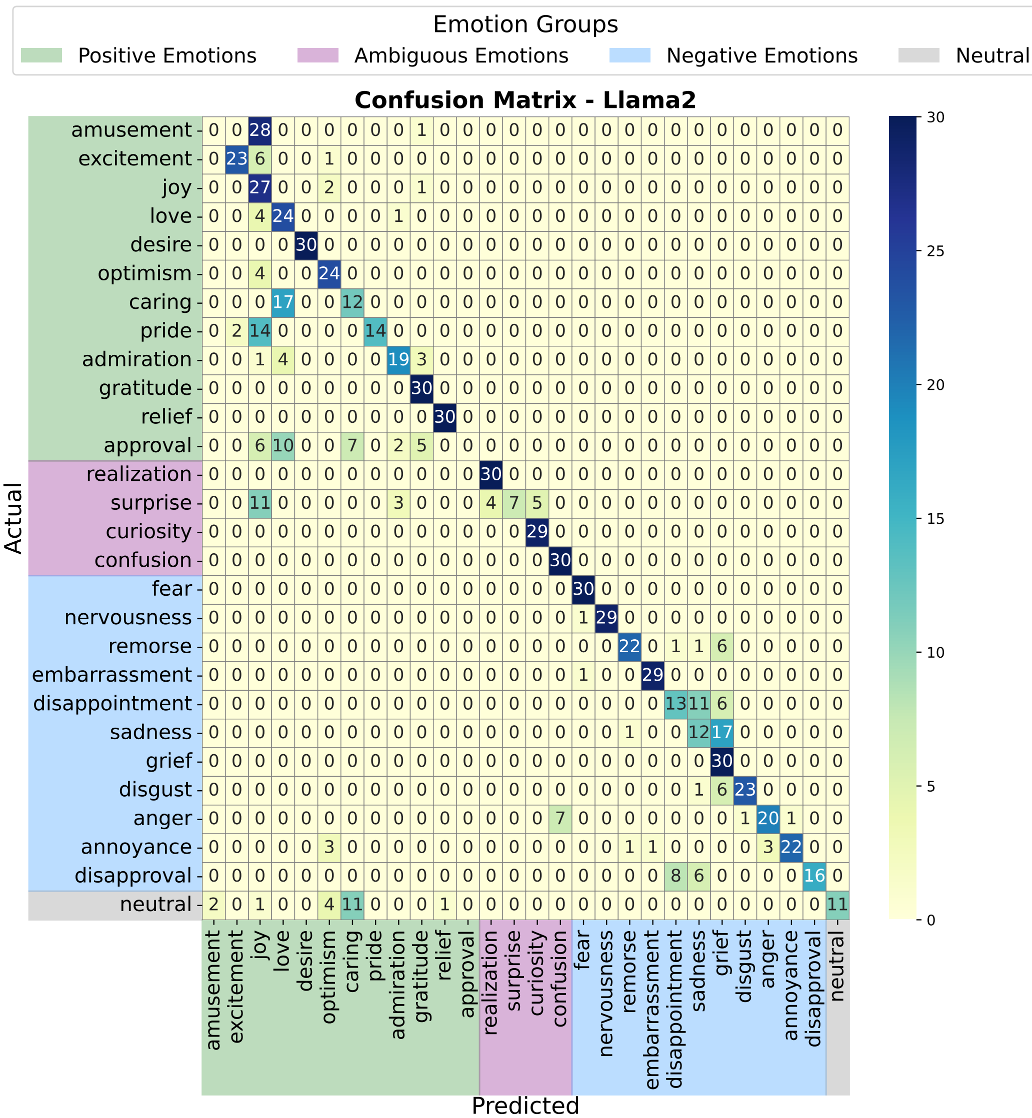


Figure 16: Llama2: Confusion matrix of emotion signals

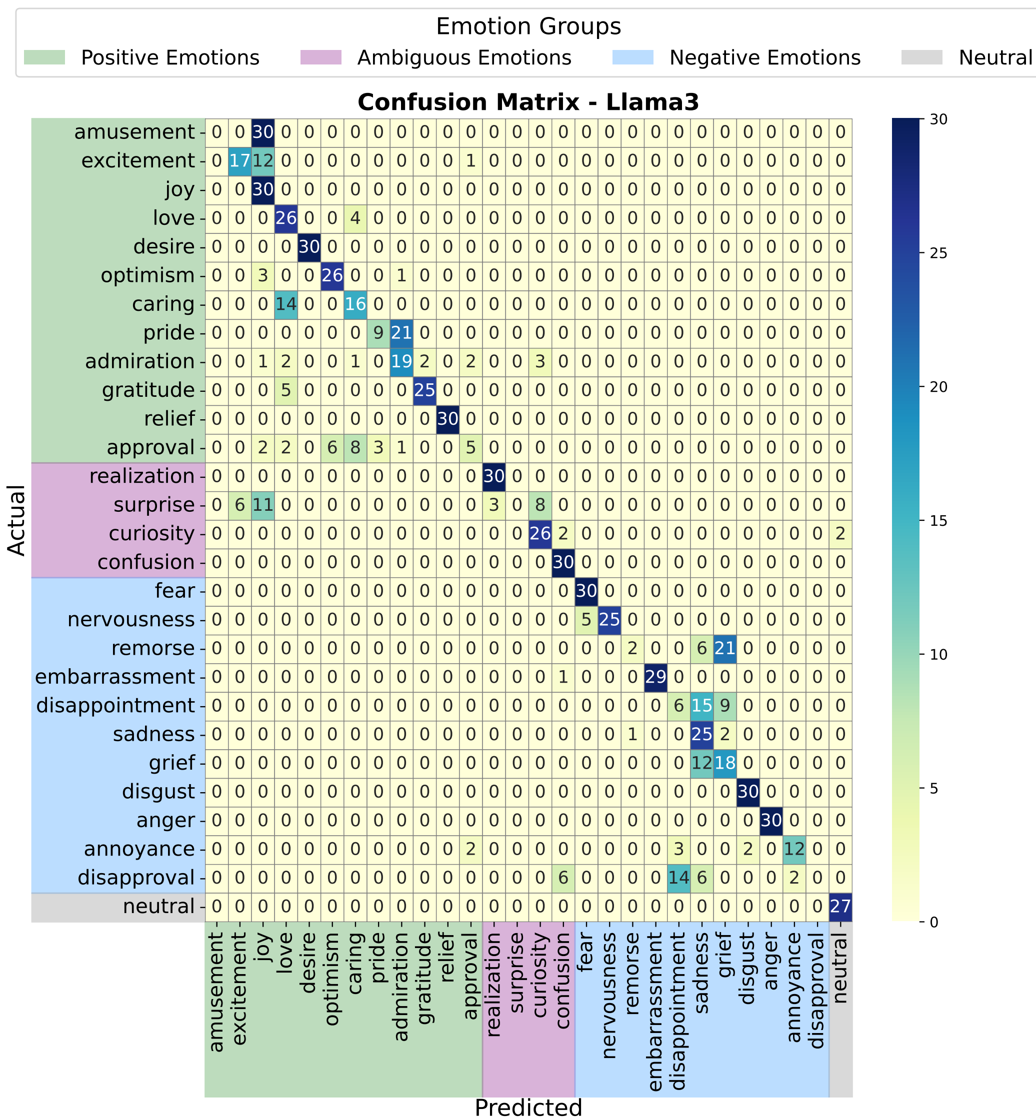


Figure 17: Llama3: Confusion matrix of emotion signals