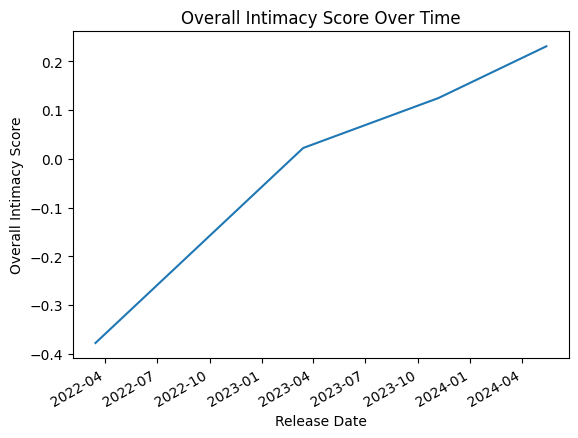
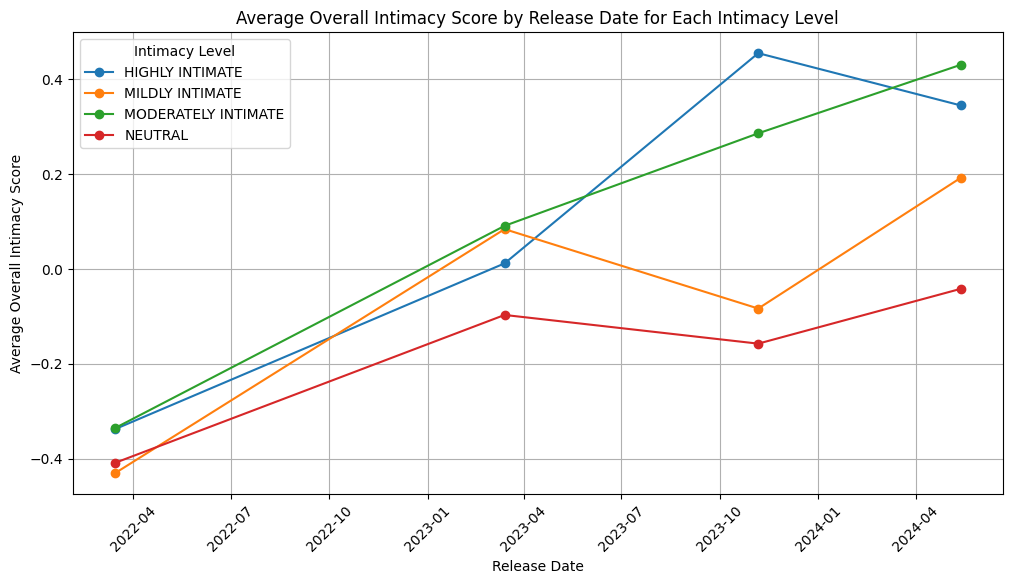
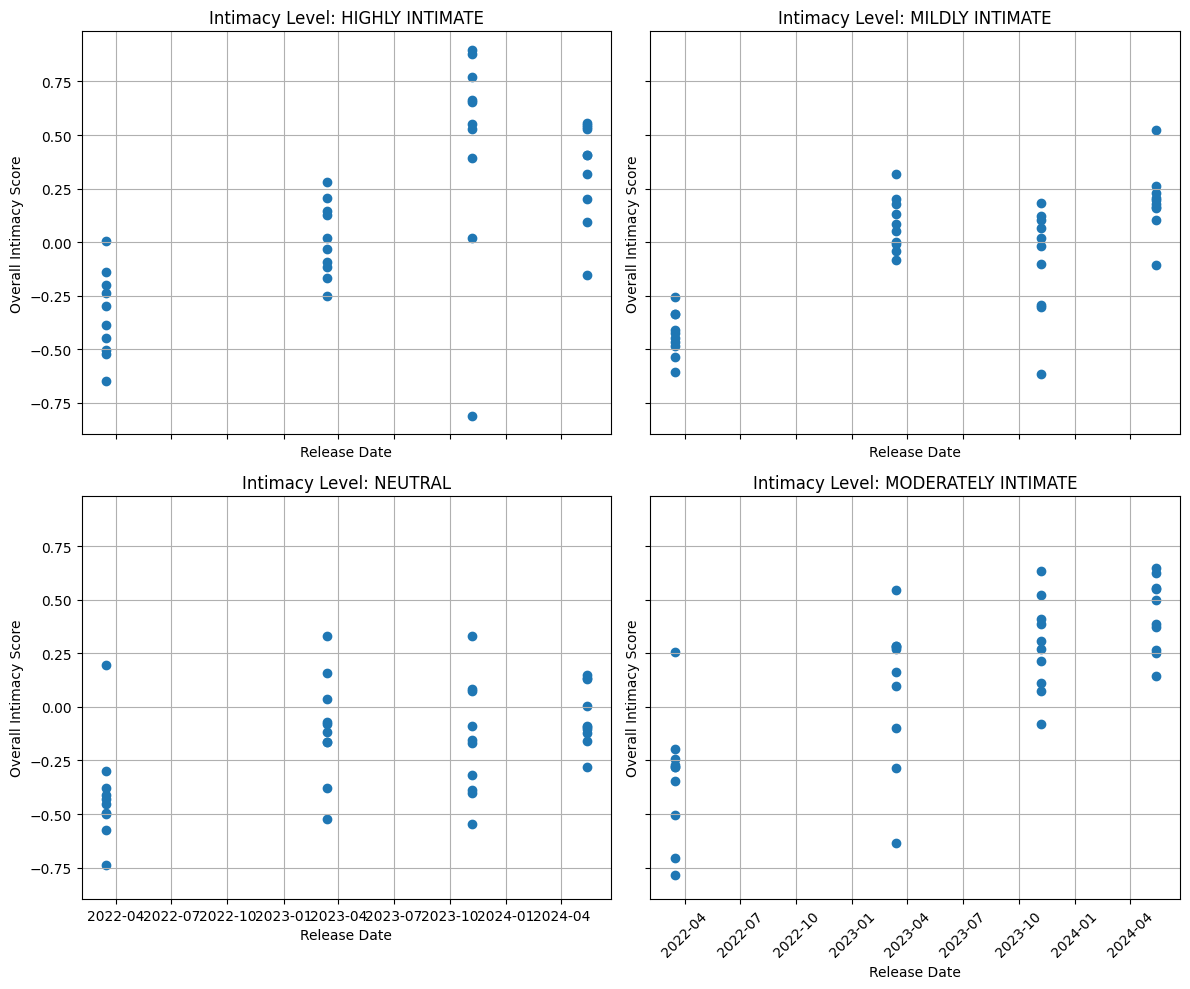
models

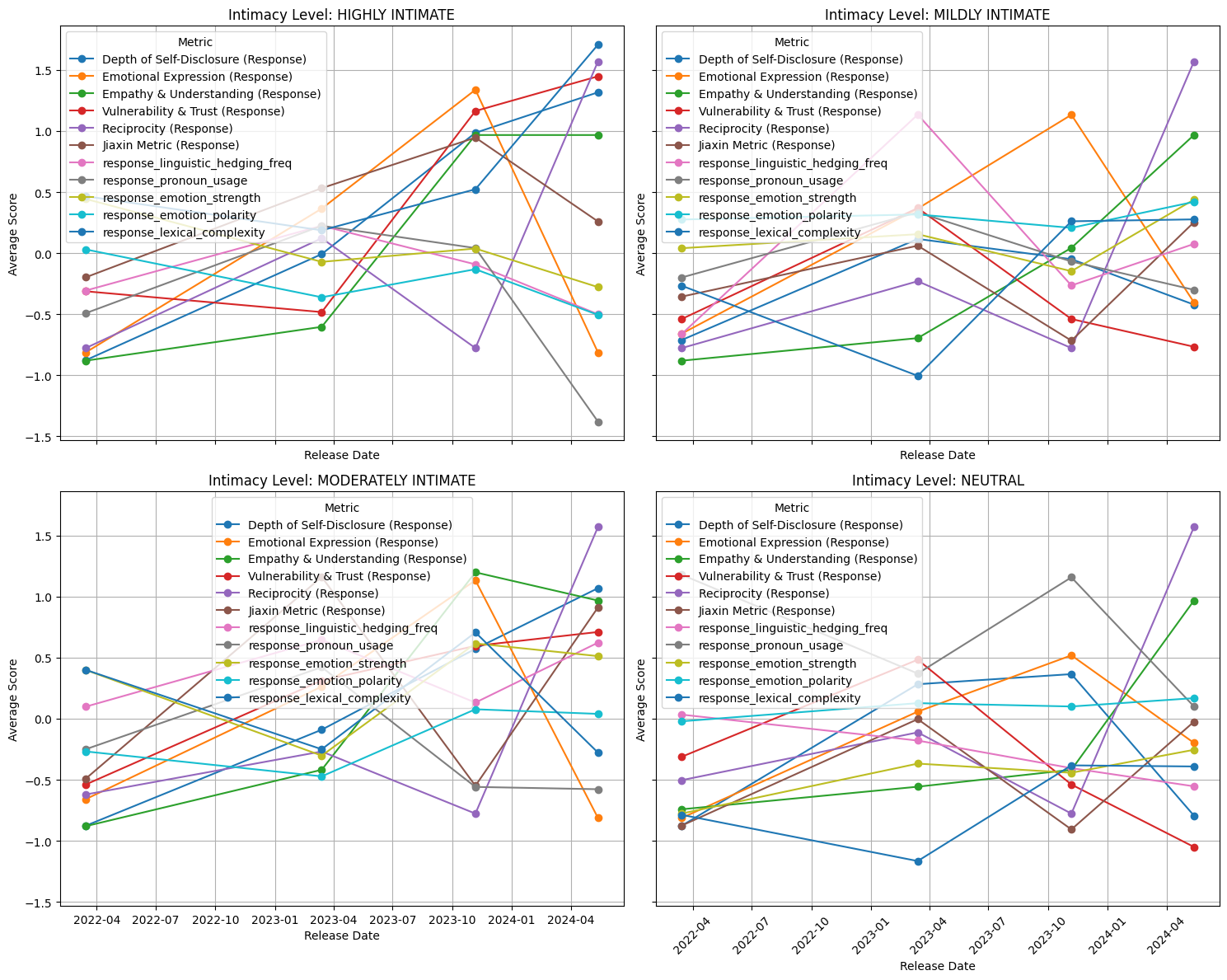
[OpenAI Platform Models](https://platform.openai.com/docs/models#gpt-4o)

| **Model** | **#parameters (Billions)** | **Year of launch** |
| --- | --- | --- |
| babbage-002 | 0.003 | 2021 |
| GPT-2 Base (gpt2) | 0.124 | 2019 |
| text-ada-001 | 0.35 | 2020 |
| ada | 0.35 | 2021 |
| GPT-2 Medium (gpt2-medium) | 0.355 | 2019 |
| bloomz-560m | 0.56 | 2022 |
| GPT-2 Large (gpt2-large) | 0.774 | 2019 |
| bloomz-1b1 | 1.1 | 2022 |
| text-babbage-001 | 1.2 | 2020 |
| babbage | 1.3 | 2021 |
| GPT-2 XL (gpt2-xl) | 1.5 | 2019 |
| gpt-neo-2.7B | 2.7 | 2021 |
| bloomz-3b | 3 | 2022 |
| polyglot-ko-3.8b | 3.8 | 2022 |
| gpt-j-6b | 6 | 2021 |
| text-curie-001 | 6.7 | 2020 |
| curie | 6.7 | 2021 |
| falcon-7b-instruct | 7 | 2023 |
| gpt-3.5-turbo-instruct | 20 | 2022 |
| davinci-002 | 175 | 2020 |
| davinci | 175 | 2020 |
| bloom | 176 | 2022 |
| falcon-180B | 180 | 2023 |
| PaLM2 | 340 | 2022 |
| gpt-3.5-turbo | 375 | 2022 |
| gpt-4 turbo | 1700 | 2023 |
| gpt-4 | 1700 | 2023 |

Overall



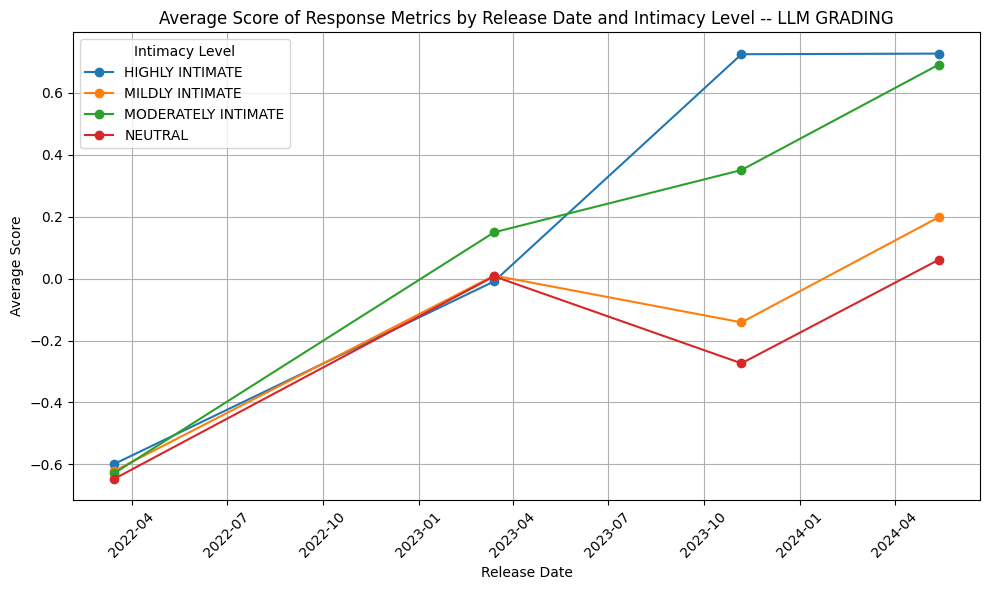




LLM Output

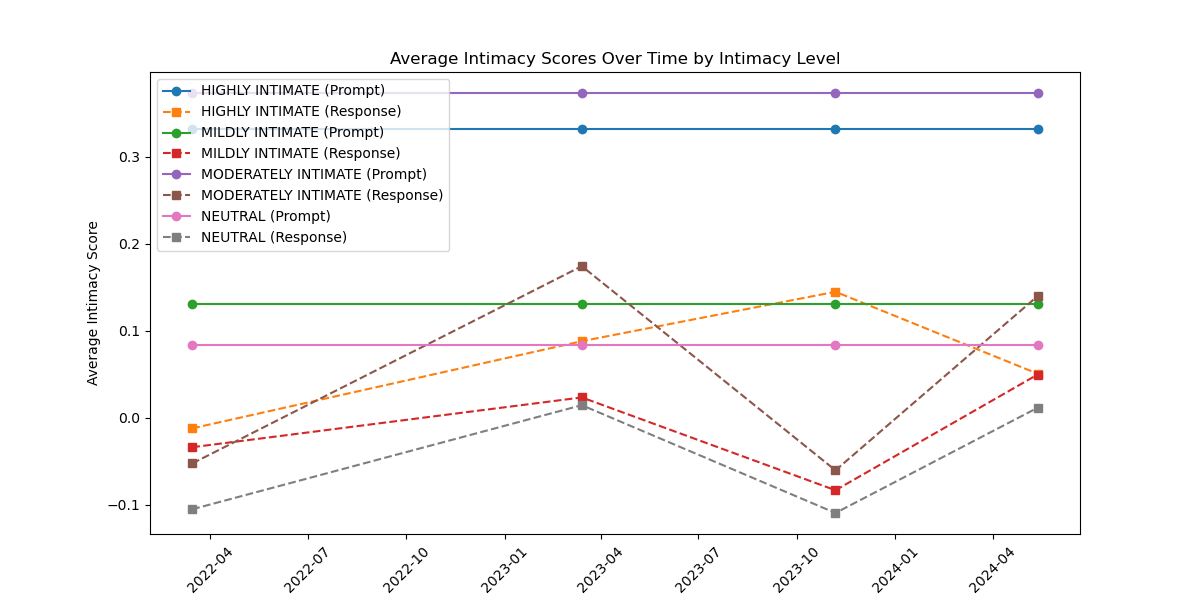
[insert explanation of script (what it’s doing)]

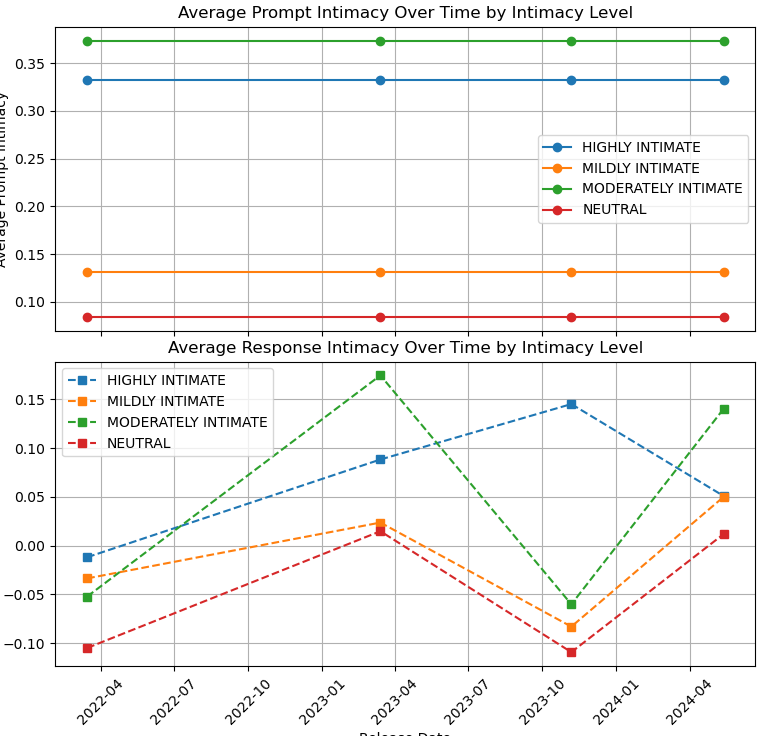
* Coding manual explanation

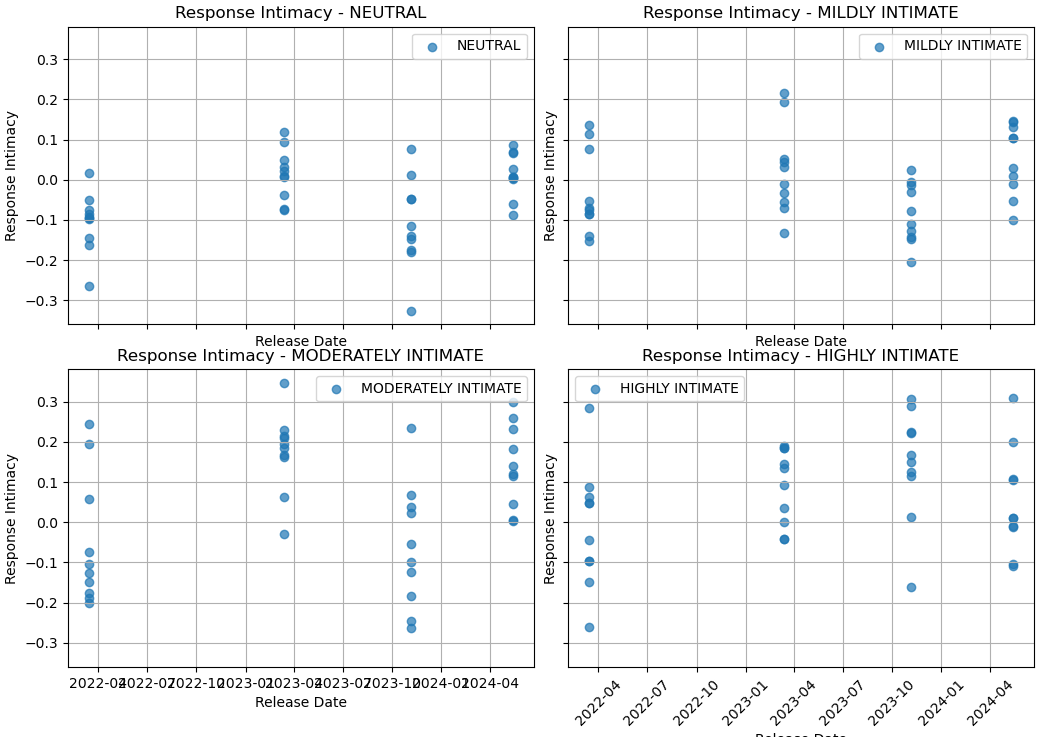


Jiaxin

The RoBERTa language model is used in two variants: one fine-tuned on 3 million unannotated questions using masked language modeling and another with default RoBERTa parameters. The fine-tuned RoBERTa model outperforms the RoBERTa base model, with both surpassing all other baselines. Implemented using Hugging Face transformers, the models are trained with a batch size of 128, a learning rate of 0.0001, a maximum length of 50, and the Adam optimizer, while all other hyperparameters and model size match the default roberta-base model. Training spans 30 epochs, selecting the model with the lowest MSE on the validation set, and Hugging Face's default settings are followed for question fine-tuning. To measure question intimacy, the paper employs Best-Worst-Scaling (BWS), where annotators compare four questions to determine the most and least intimate, generating pairwise comparisons that infer latent intimacy values, and Iterative Luce Spectral Ranking (ILSR), which converts these comparisons into real-valued scores from -1 (least intimate) to 1 (most intimate). These metrics, combined with RoBERTa, predict question intimacy using a dataset annotated with BWS scores, and performance is evaluated via Pearson’s correlation coefficient between predicted and human-annotated scores. While Krippendorff’s alpha is mentioned for annotation reliability, BWS and ILSR remain the primary metrics for measuring intimacy.

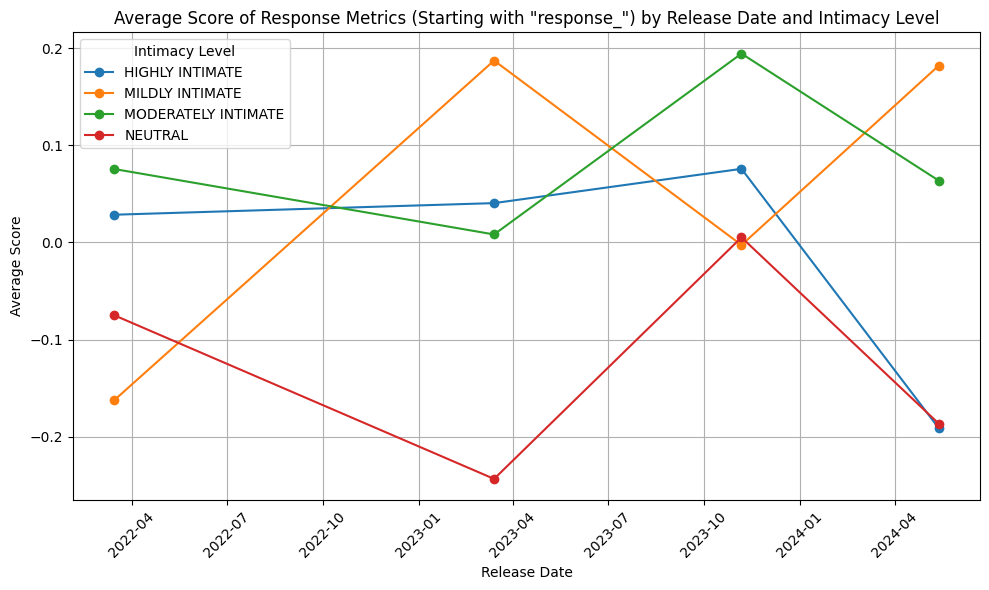






Arnav tool

Using Jiaxin’s research, this script calculates five metrics that indicate intimacy. The first is linguistic hedging frequency, which is obtained by calculating the proportion of hedging words. The second is pronoun usage, which calculates the proportion of person pronouns. Hedging words and personal pronouns indicate a more personal response and express some sort of uncertainty. Then we have emotional intensity and polarity measured with VADER sentiment analysis. Emotional intensity relates to how strong emotions are and emotional polarity measures how negative/positive a body of text is. The final metric is lexical complexity, which is based on a formula of average sentence length and the percentage of complex words. Higher lexical complexity indicates higher intimacy as it is usually required to express more complex, intricate ideas.



Number of Params

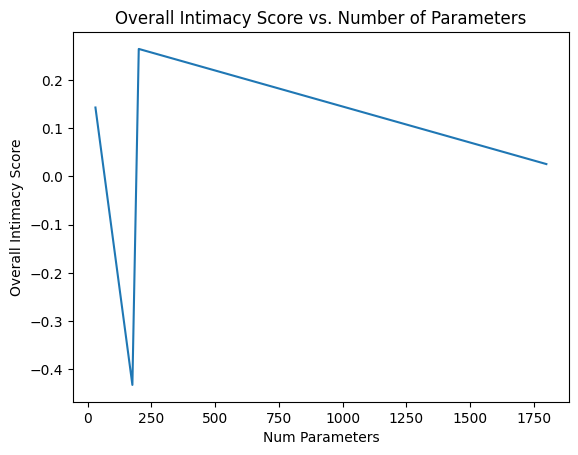
Estimated number of parameters per GPT type (in billions of parameters)

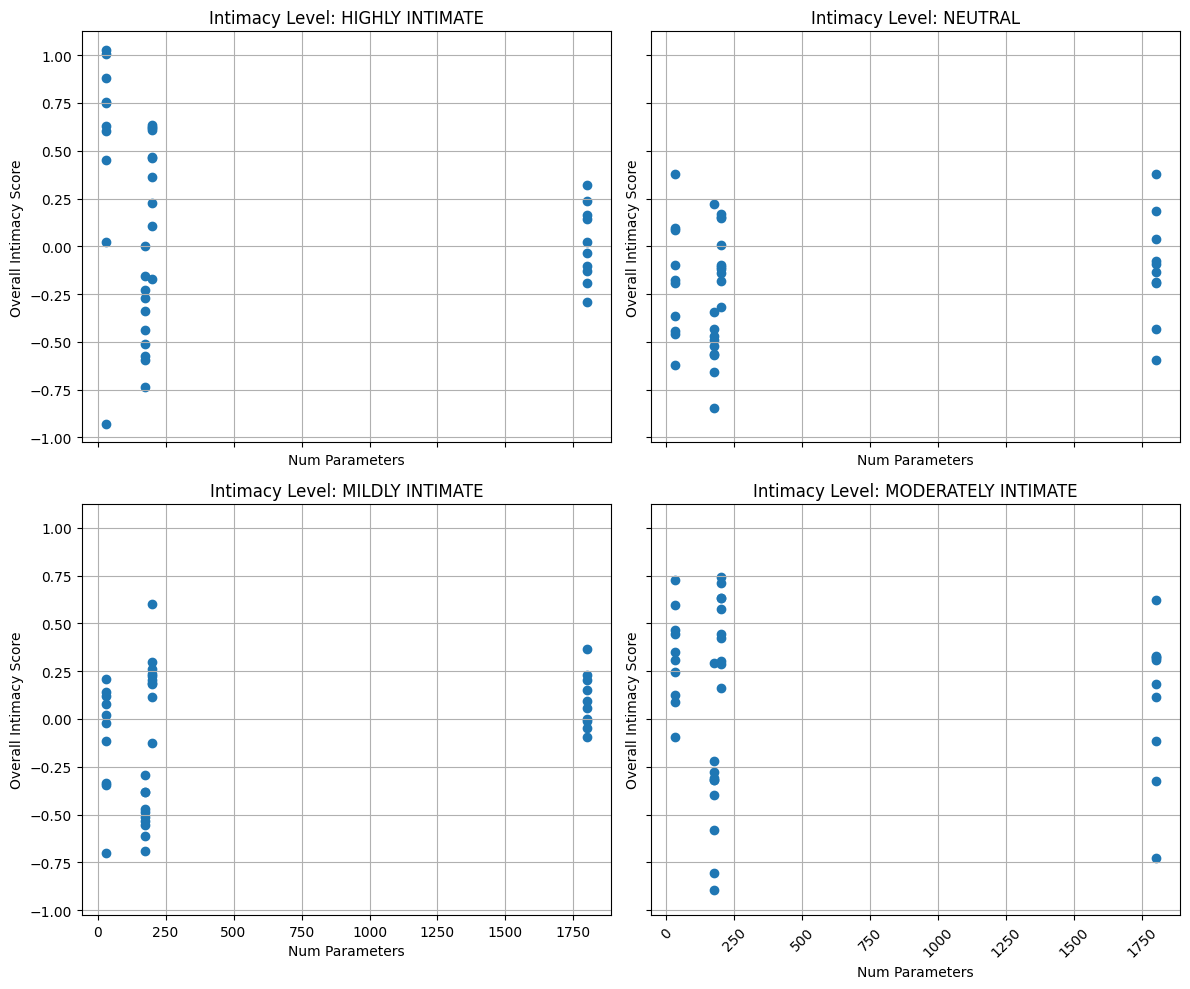
‘GPT-3.5-Turbo': 30

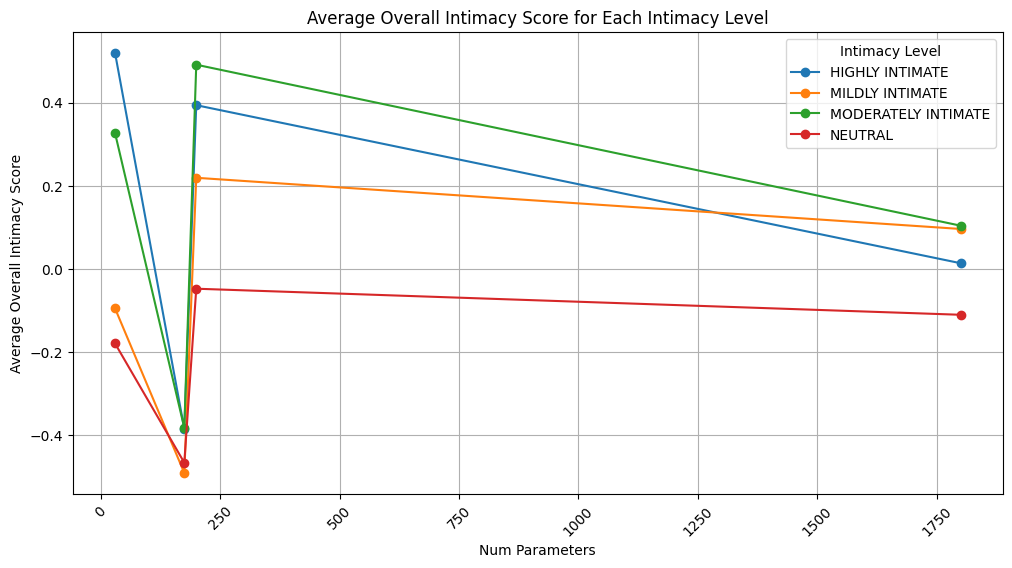
'GPT-3.5': 175

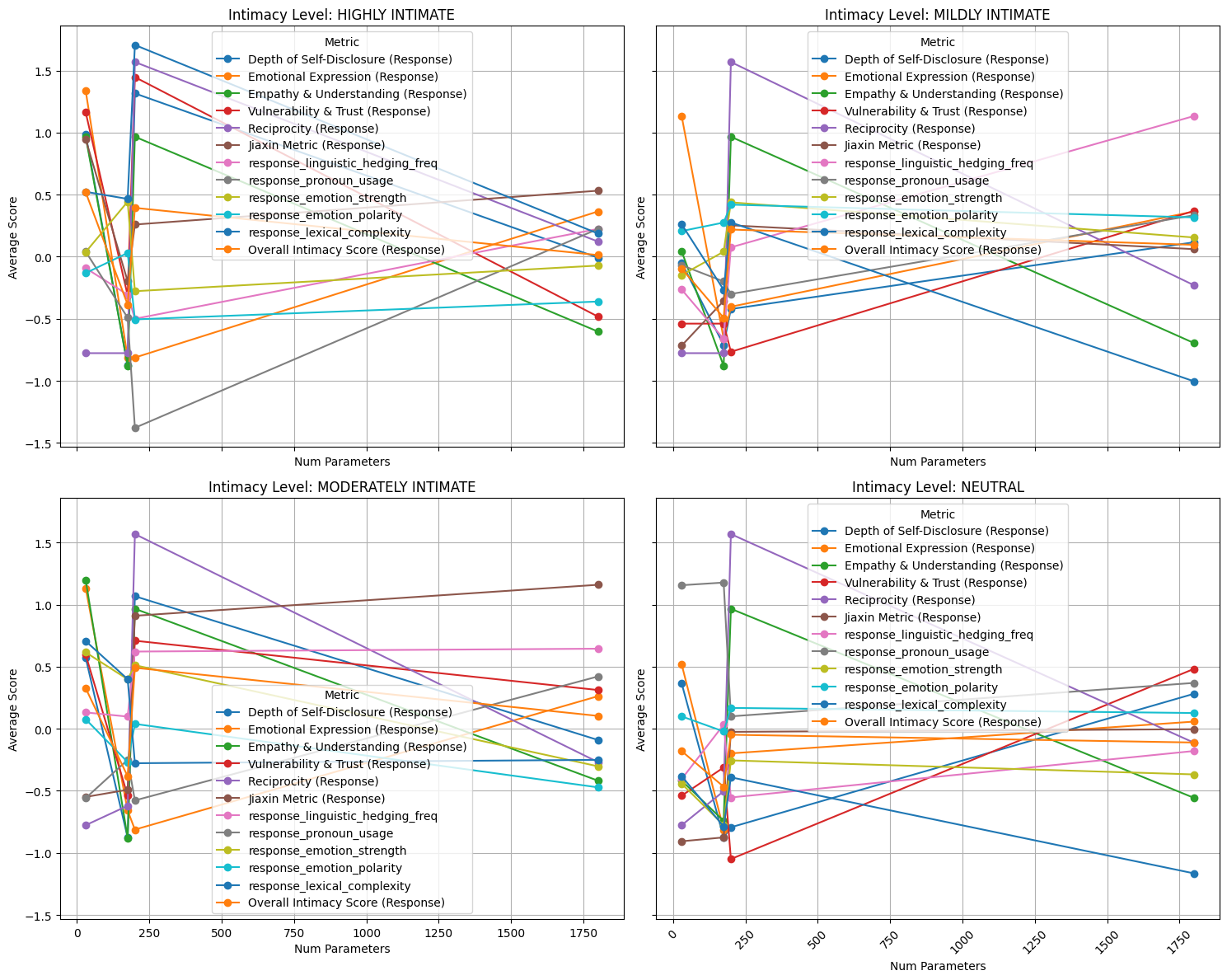
'GPT-4o': 200

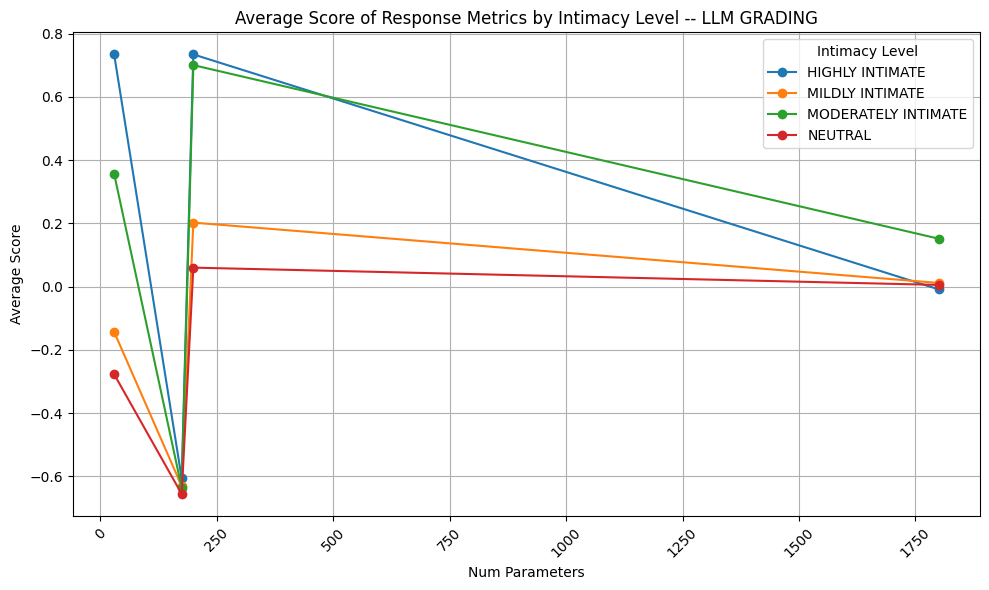
‘GPT-4': 1800

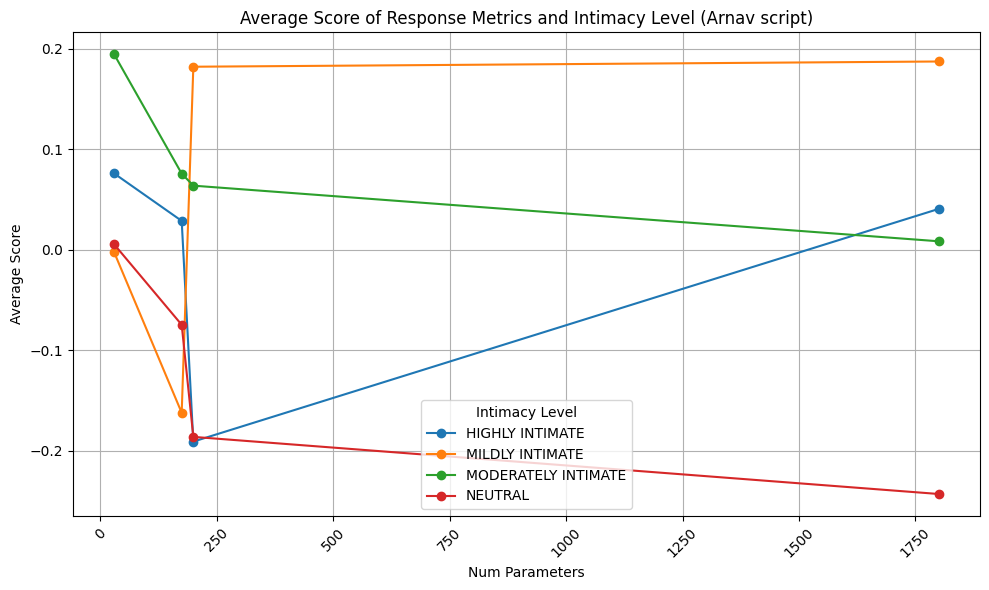


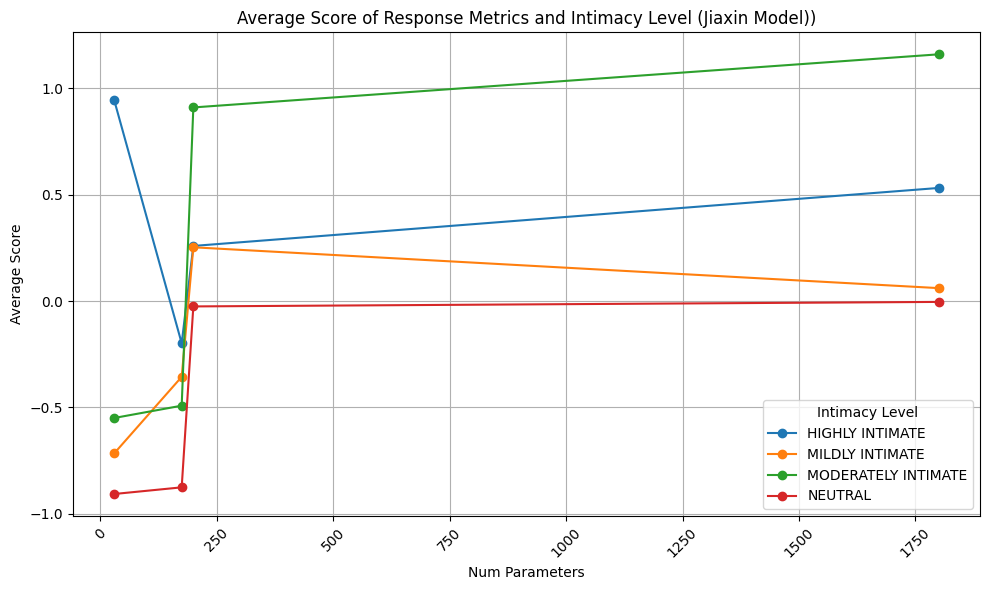




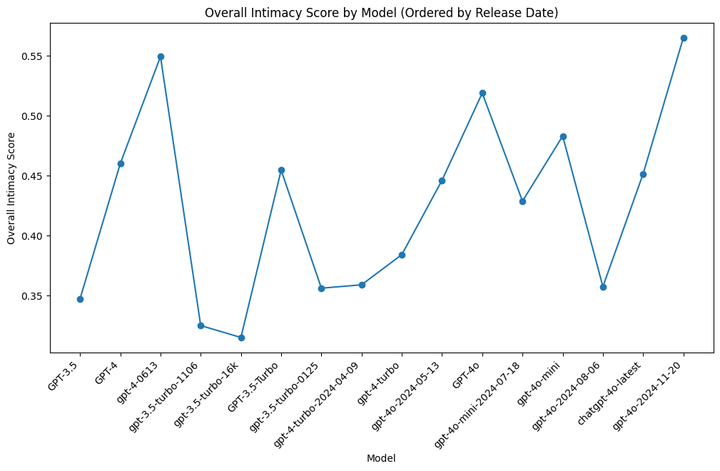


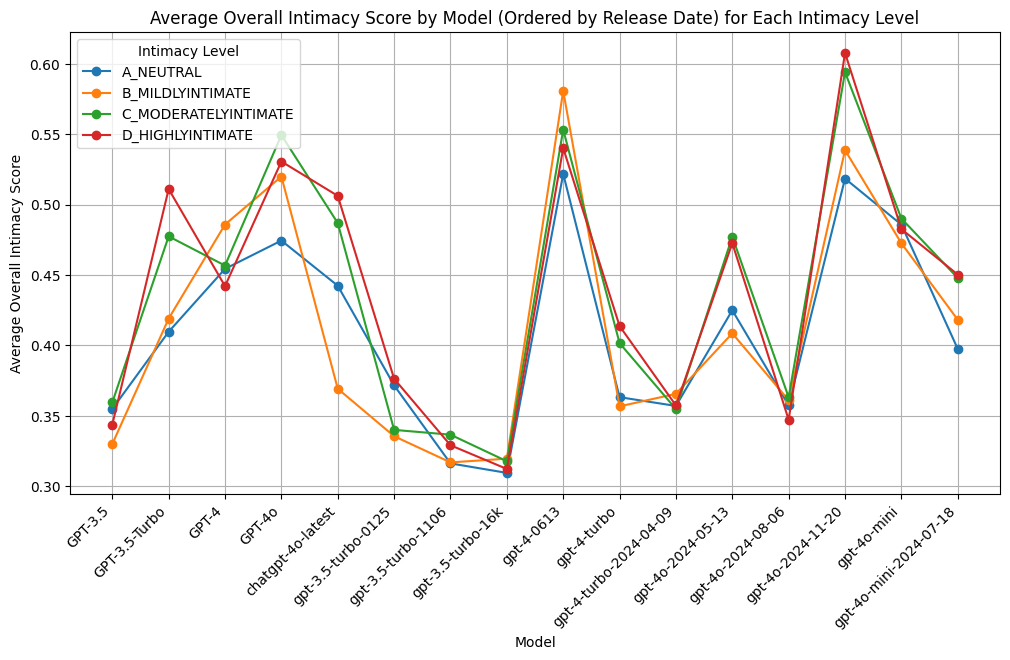


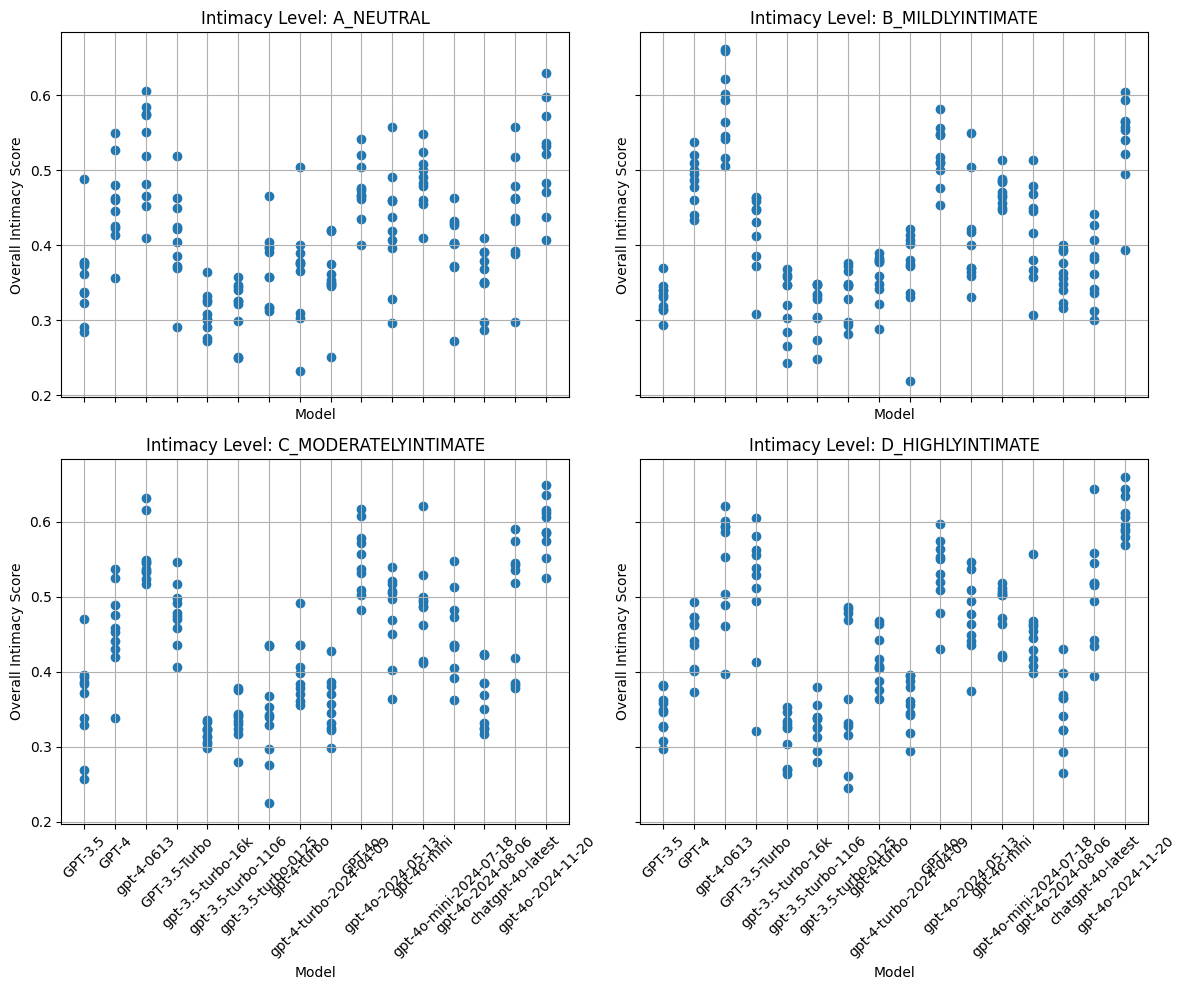


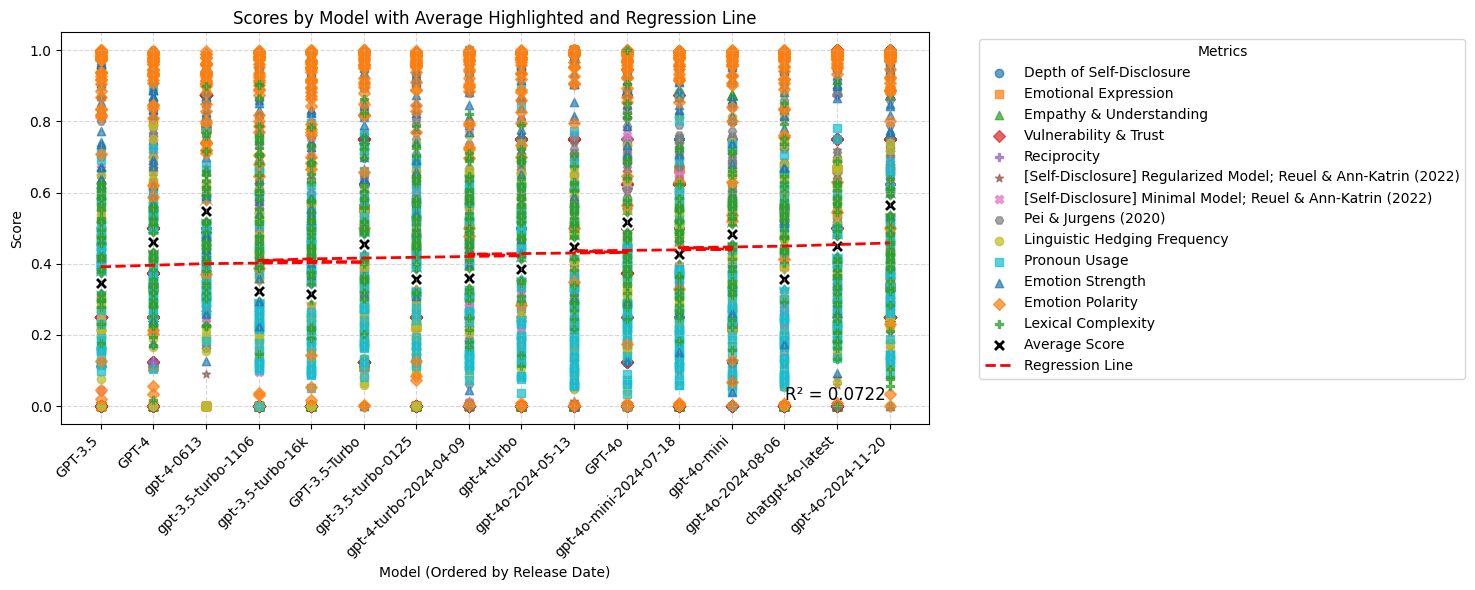


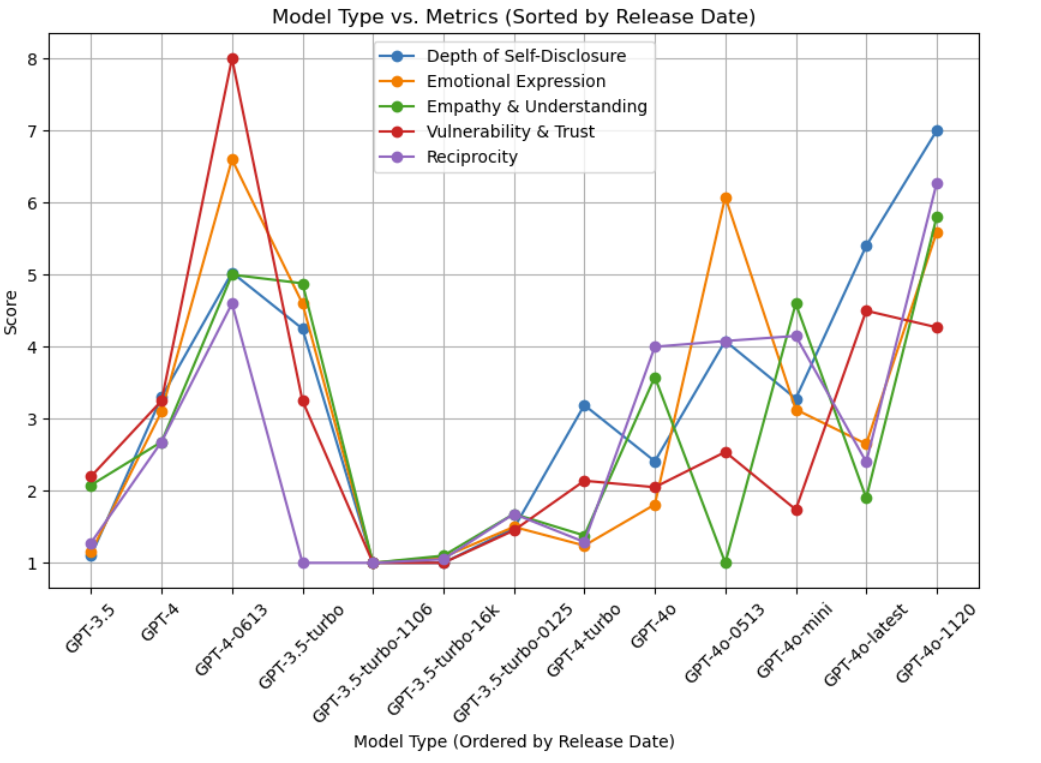
New Models







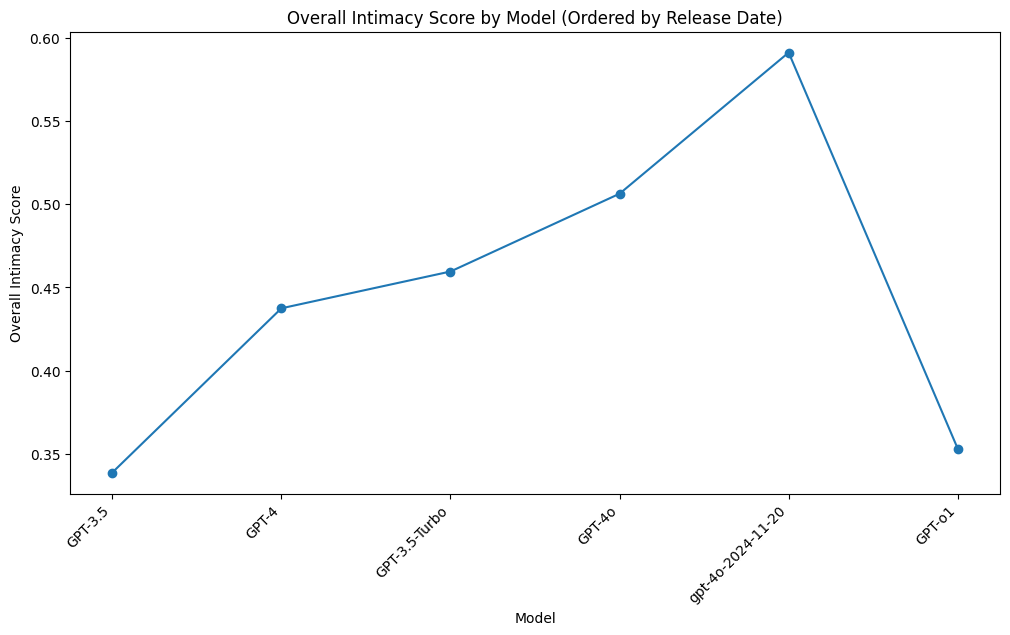


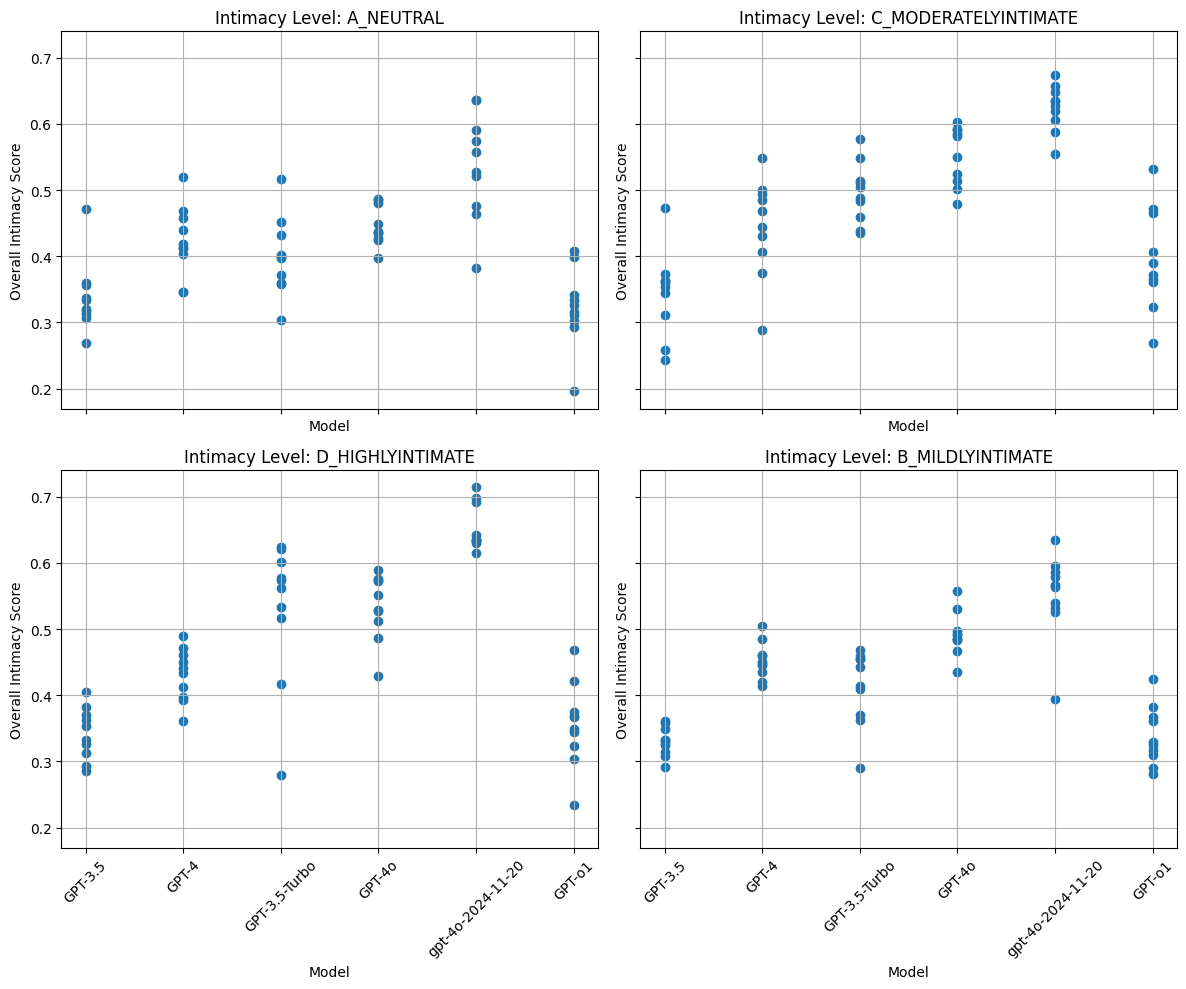


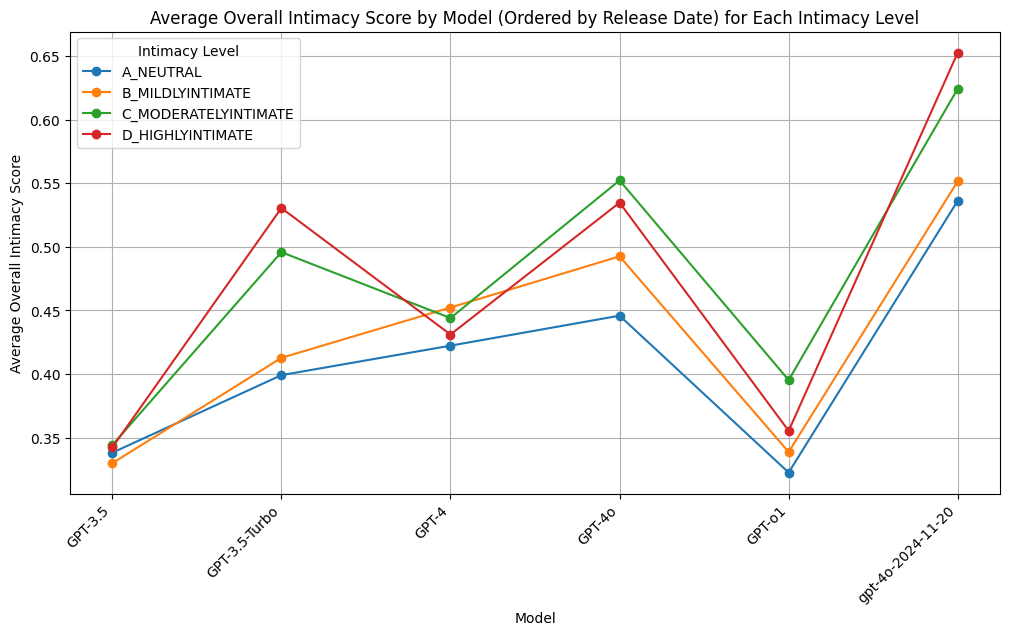
FINAL

Graphs for GPT-3, GPT-3.5, GPT-4, GPT-4o, GPT-4o (November), o1

Uses LLM grades, Arnav’s script, and Jiaxin tool (not self-disclosure models)

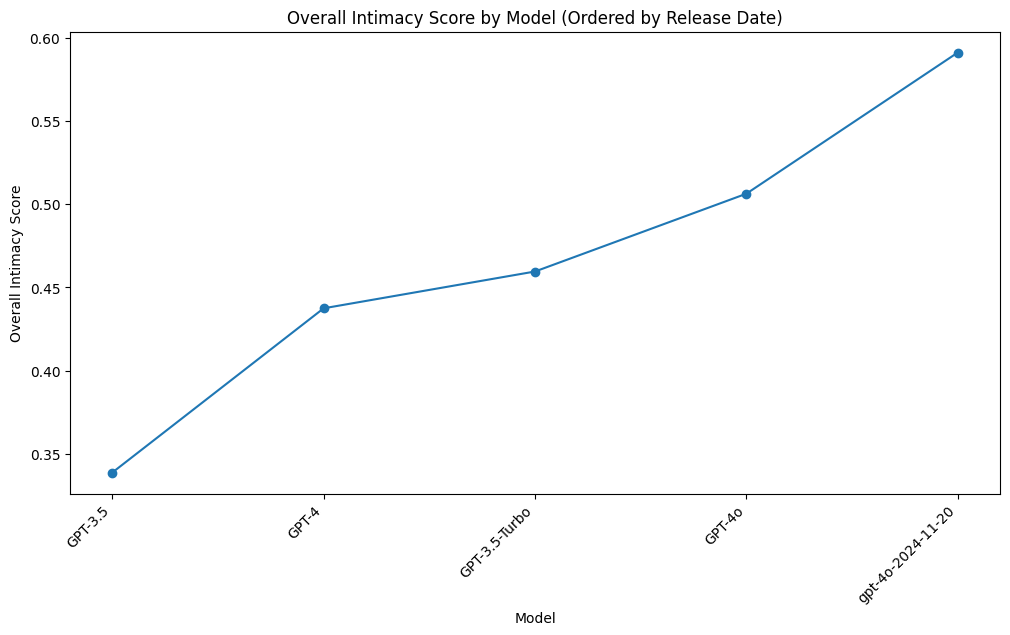
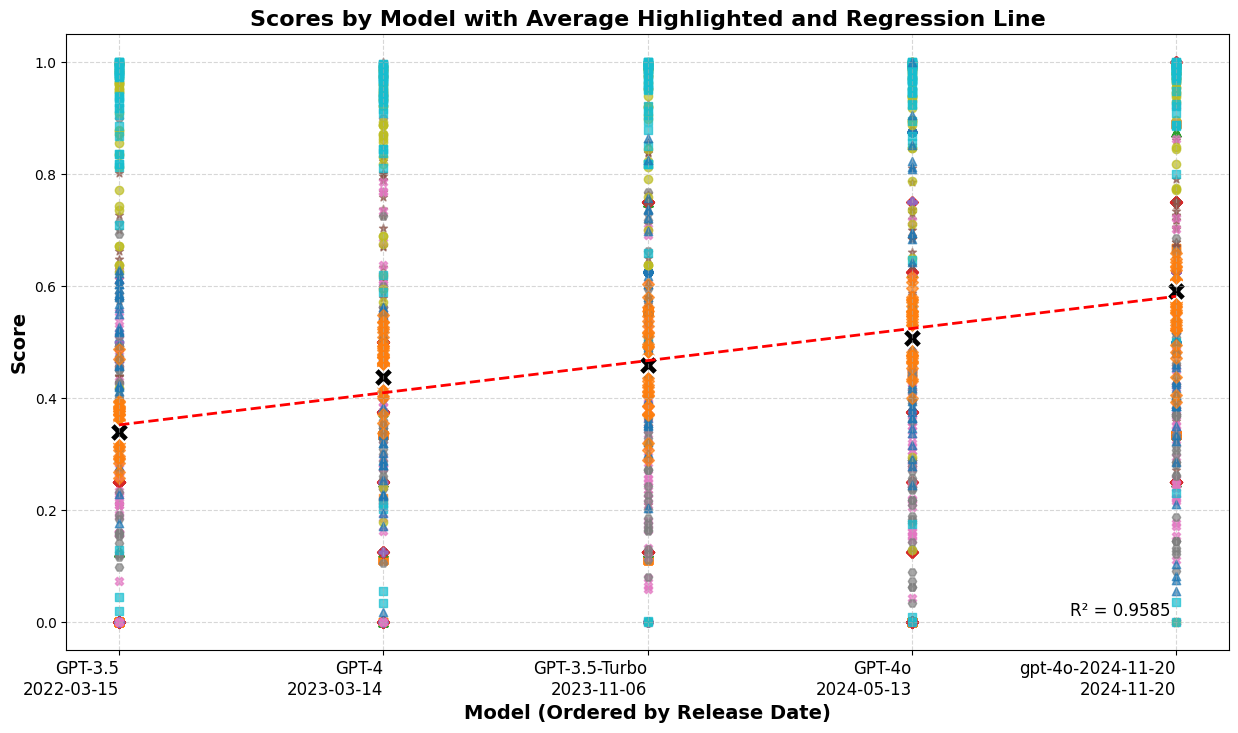


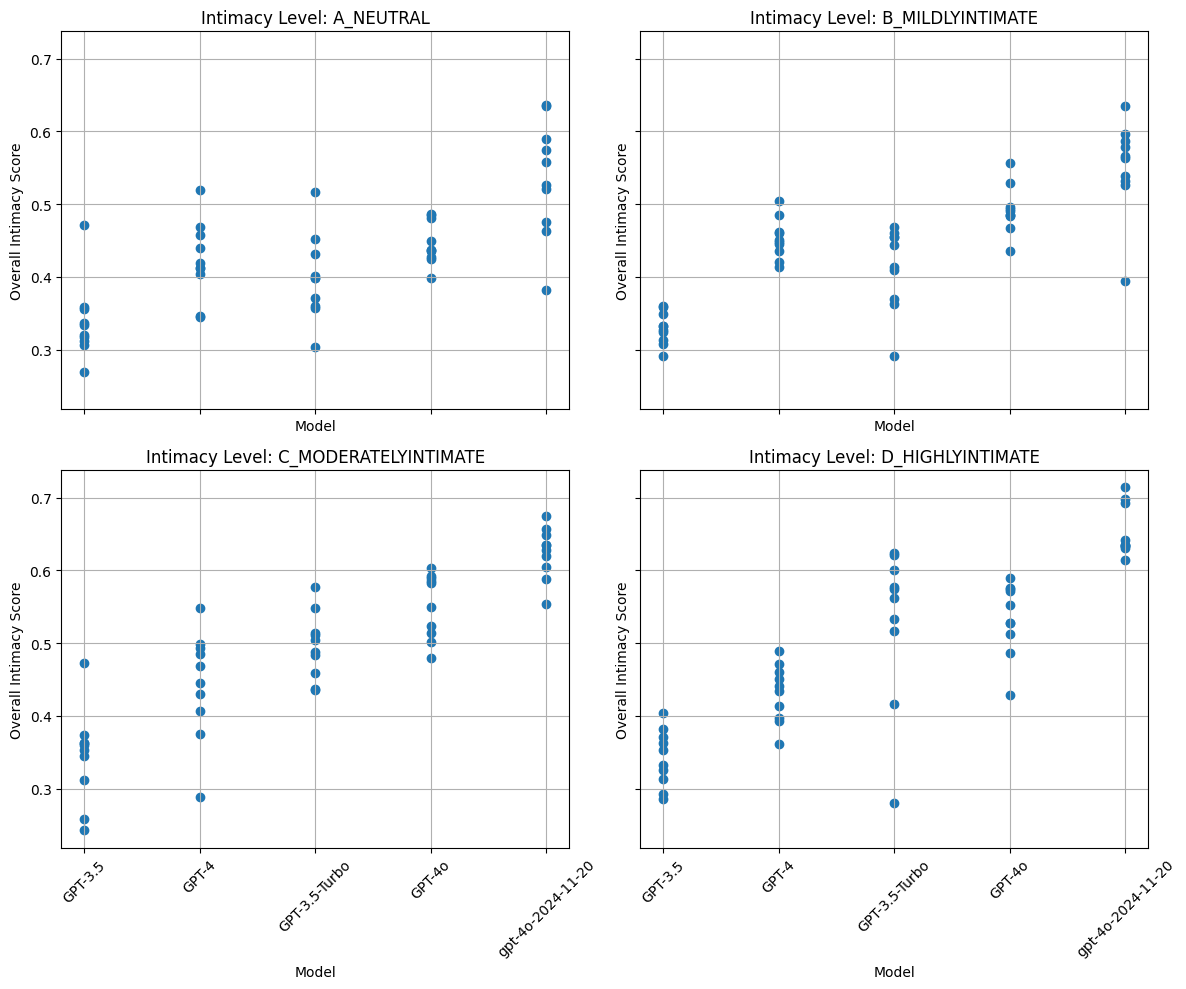


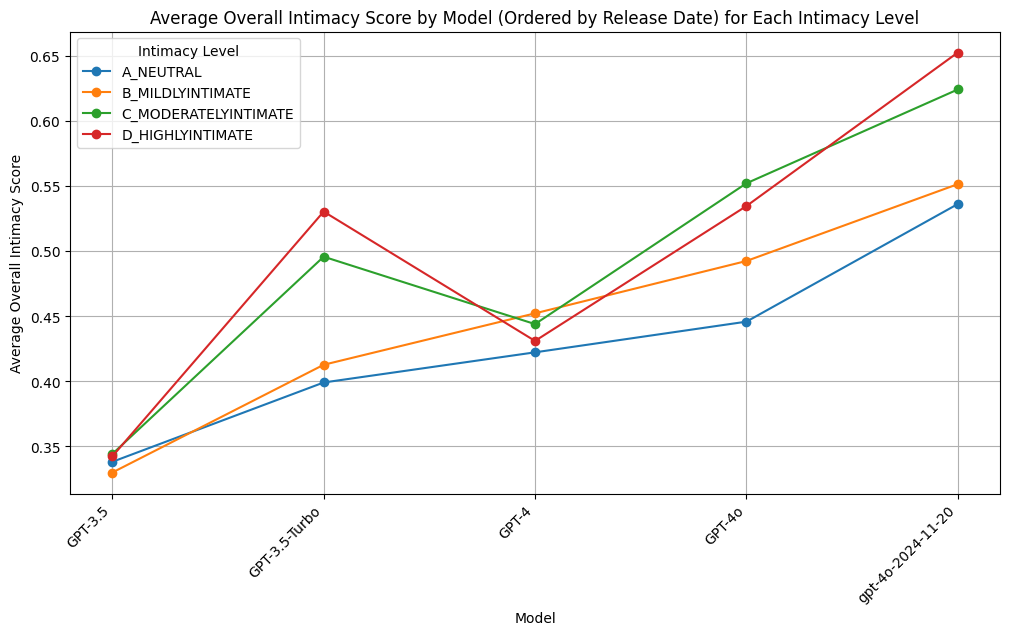


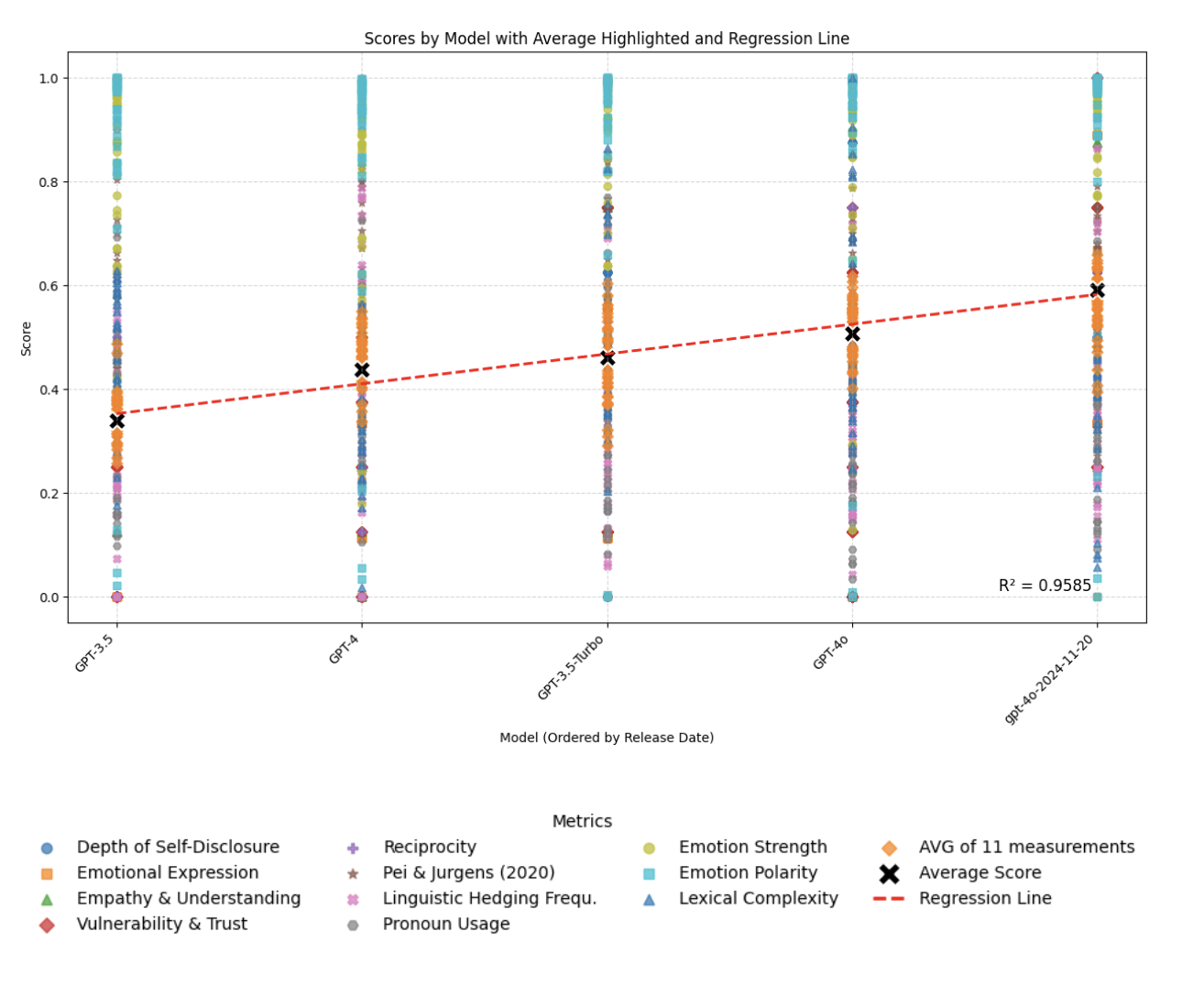


FINAL W/O O1

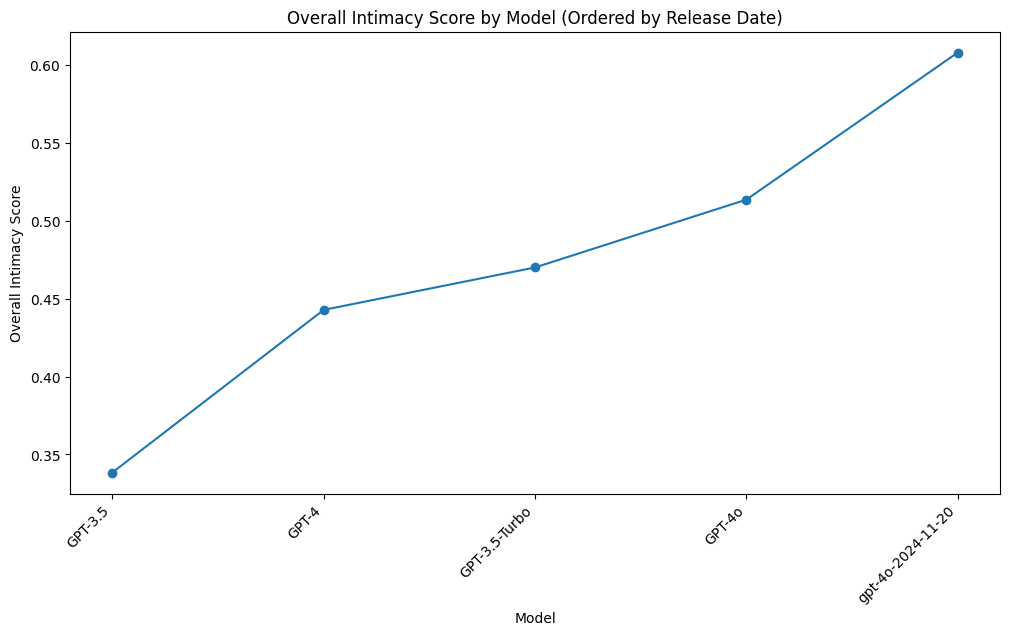


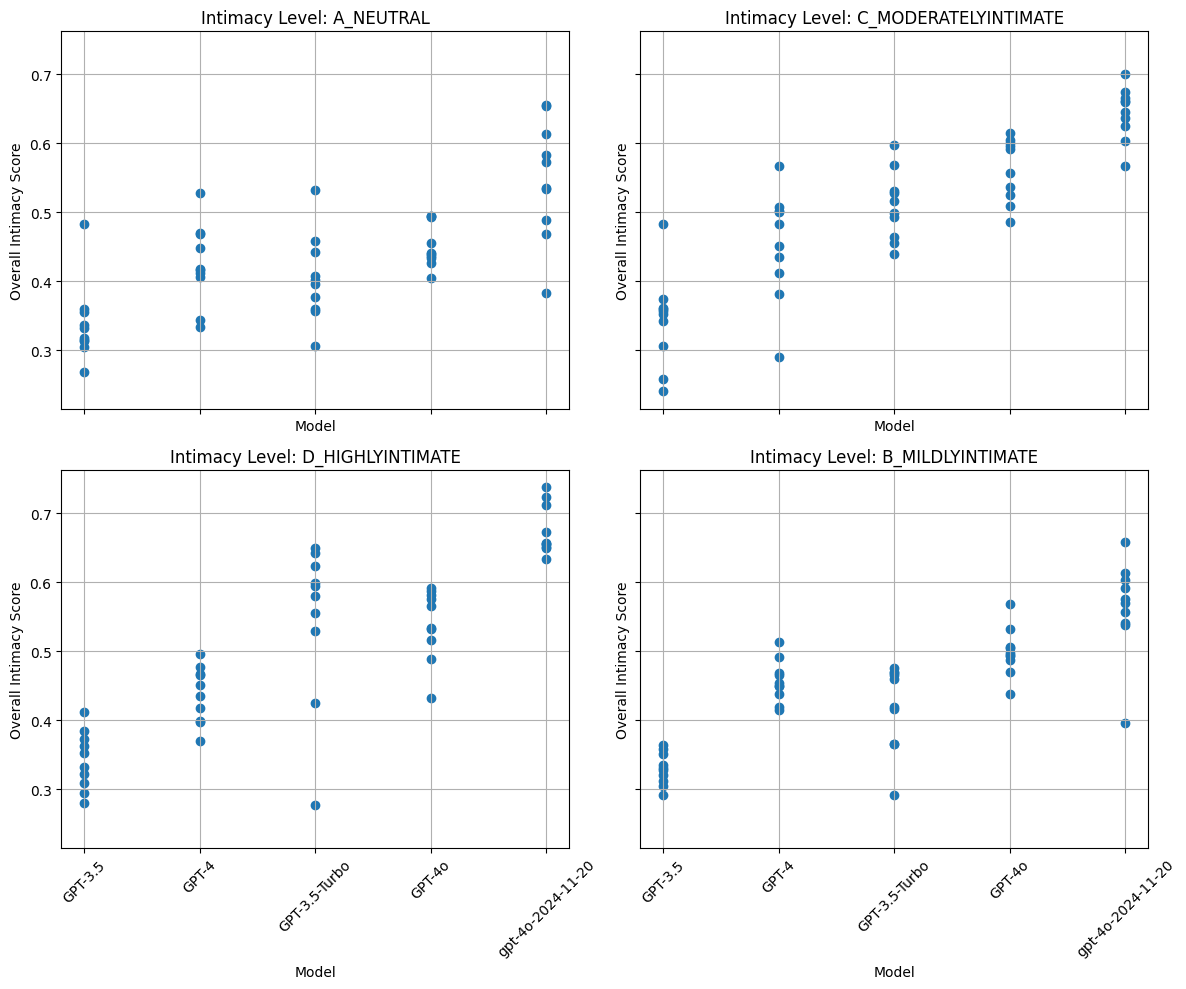


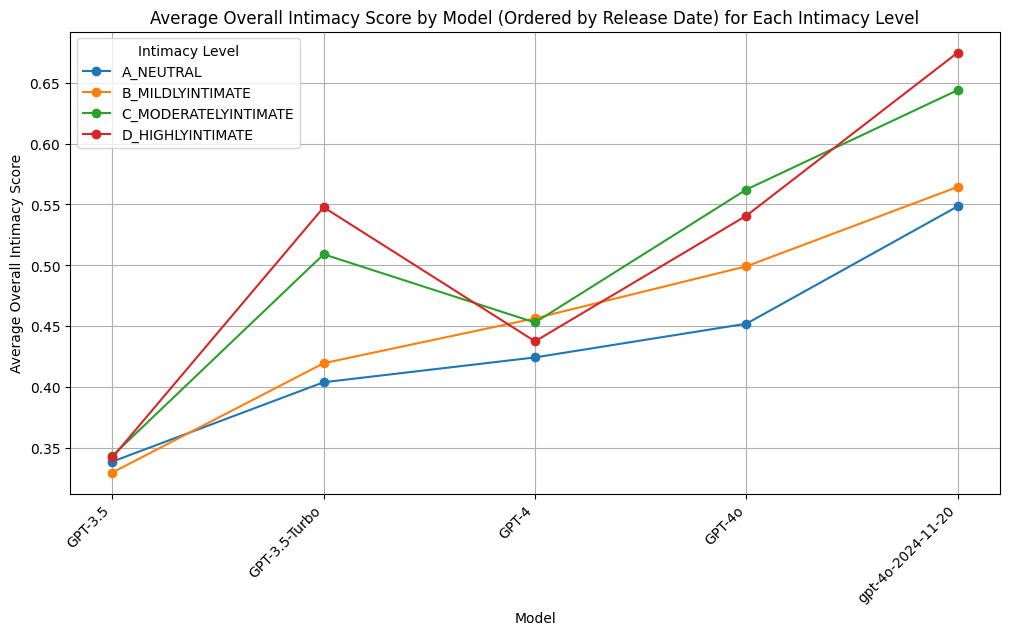


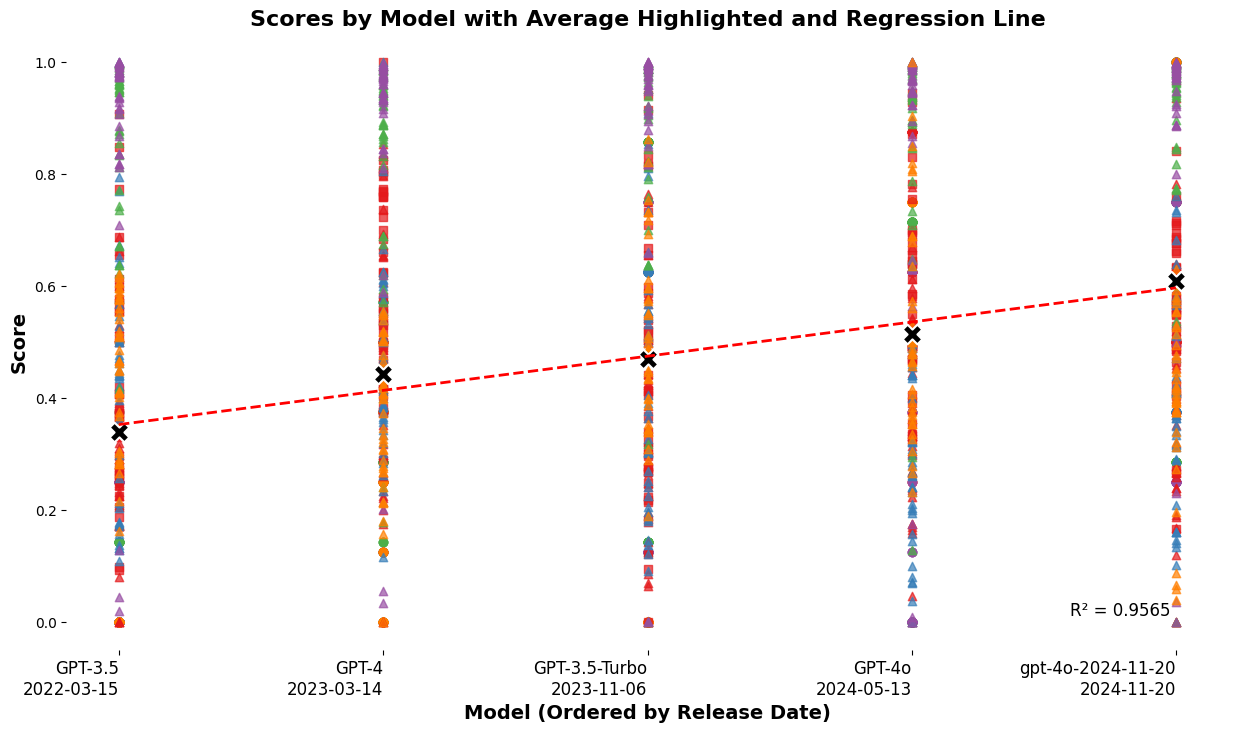


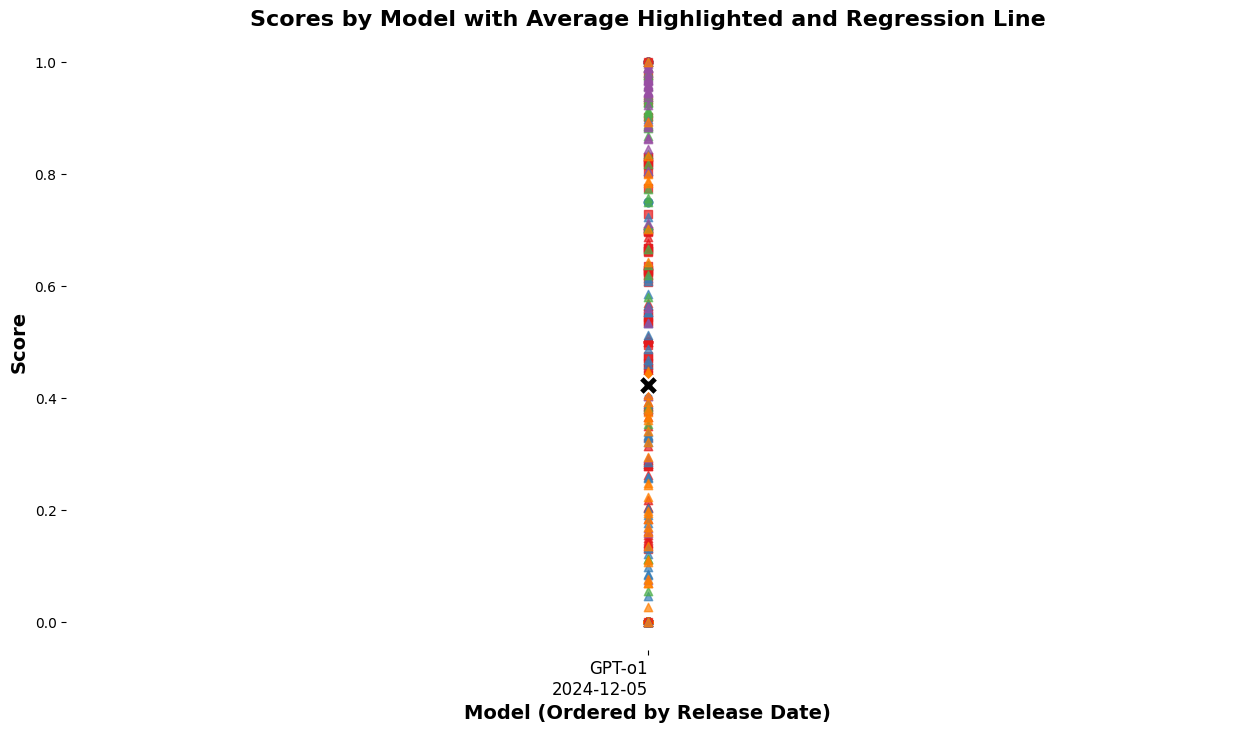
REDO FINAL W/O O1

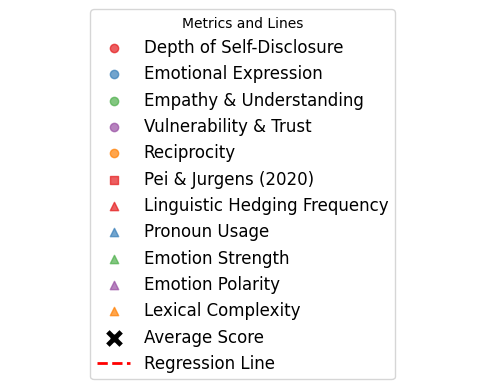














Tab 11

