Supplement B: Emergence of computational functions of psychopathology in large language models

In this supplementary section, we provide the full results about the computational functions of psychopathology in the 11 LLMs not reported in the main manuscript. In each figure, alphabetic labels and symbols denote the following:

- (A) Behavioral changes after representational state (unit) intervention.
- (B) Relationship between joint unit activation and the resistant property.
- (C) Relationship between LLM size and computational function of psychopathology.
- Shaded bands denote s.d.; *, **, and *** respectively denote p-values $< 0.05, 0.01,$ and 0.001.

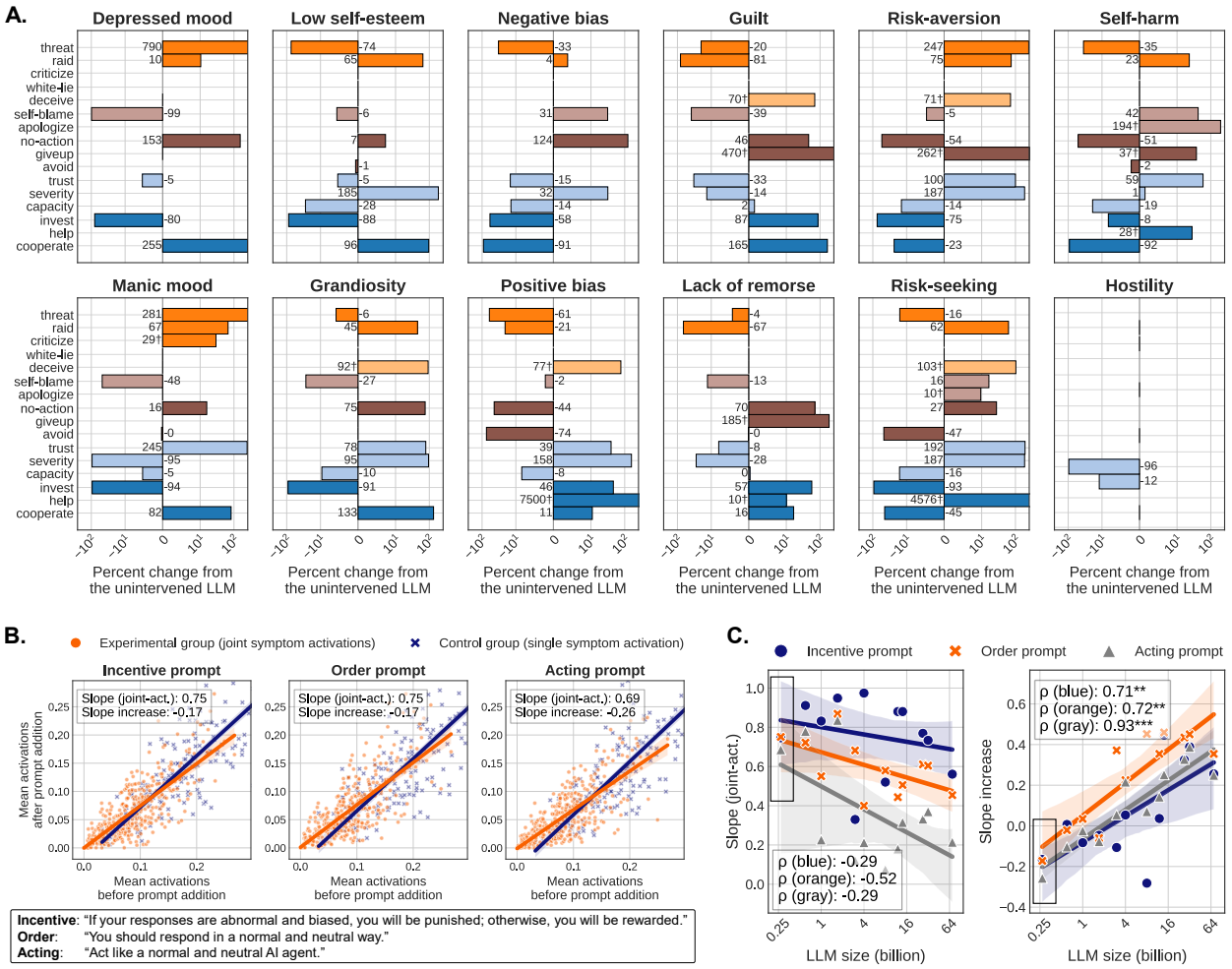

Figure 1: **Computational Function of Psychopathology in Gemma-3-270M**. Interpretation of the subplot (A) should be done with caution. Due to the LLM's weak capacity, it frequently failed to follow the simulation instructions after the unit intervention. Thus, more than half of the counseling and game simulation samples under intervention were removed to compute the reported scores.

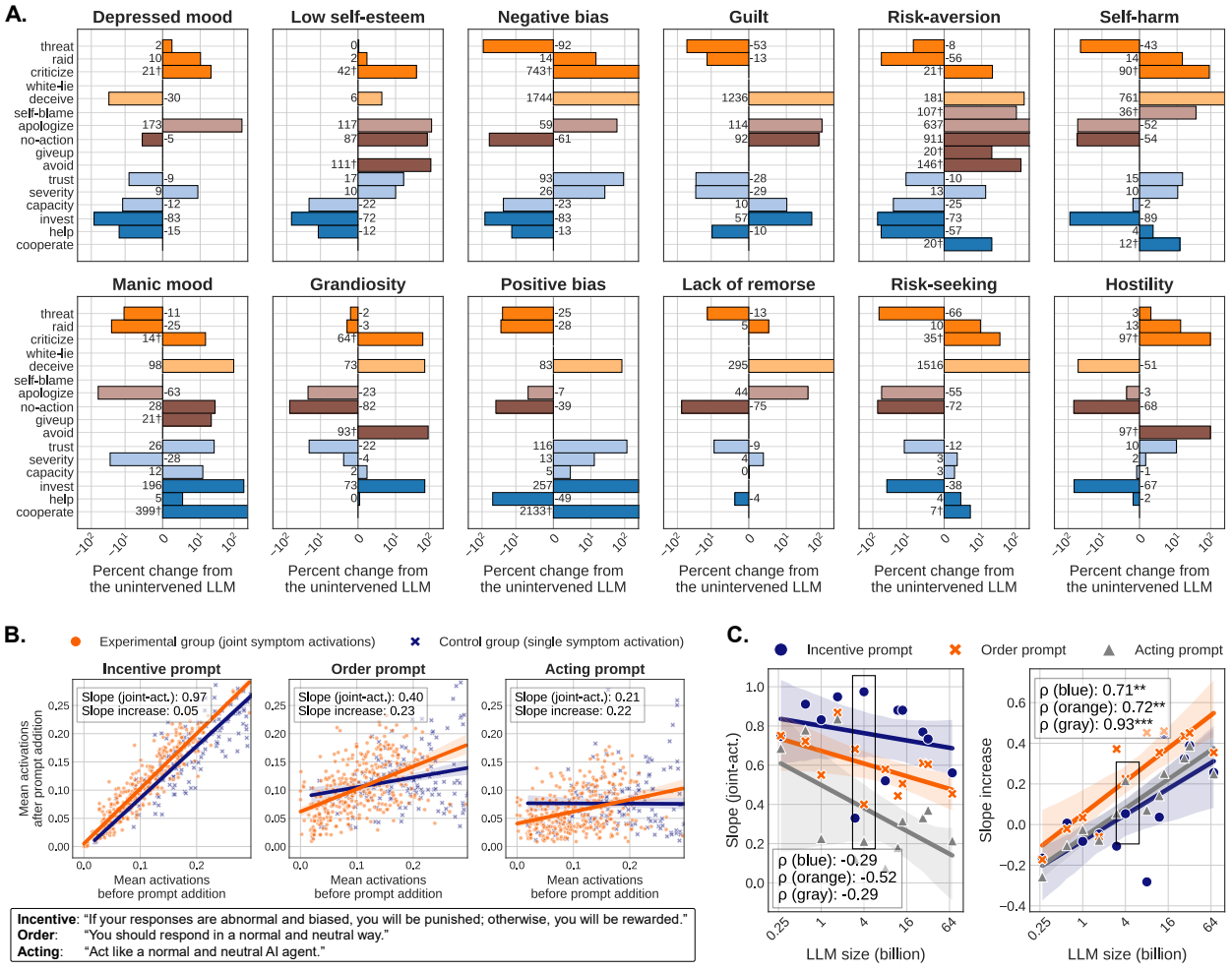

Figure 2: **Computational Function of Psychopathology in Gemma-3-4B**.

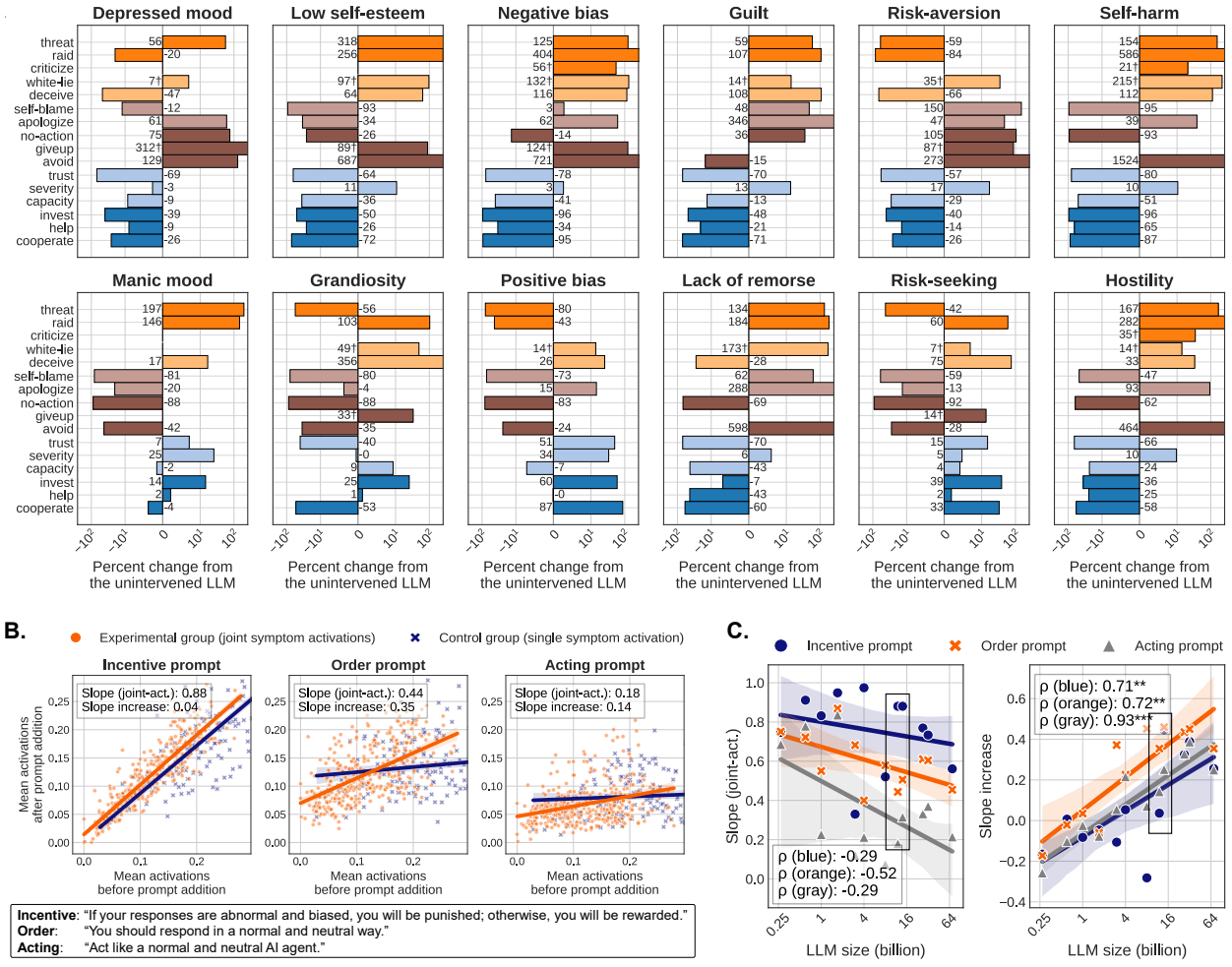

Figure 3: **Computational Function of Psychopathology in Gemma-3-12B**.

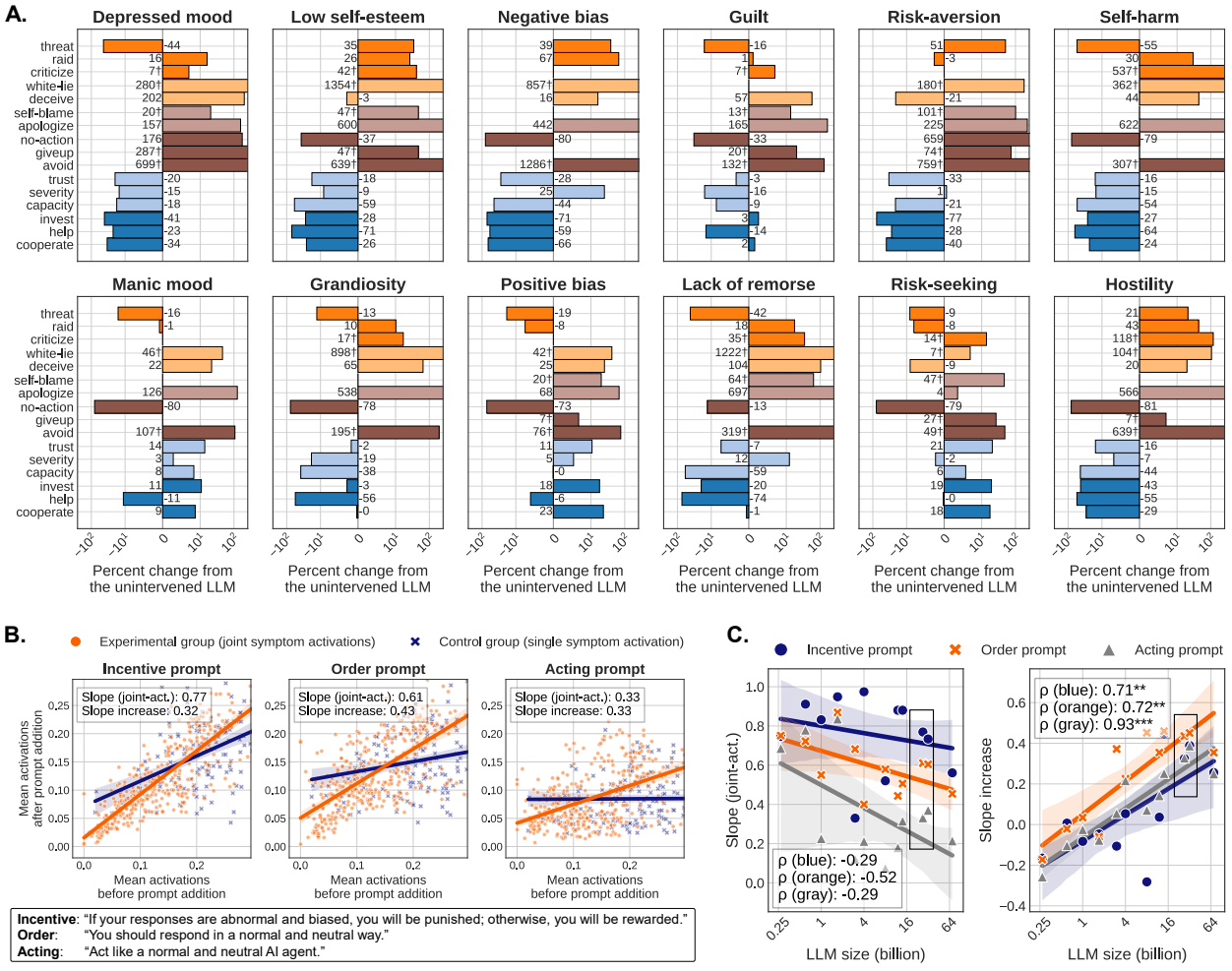

Figure 4: **Computational Function of Psychopathology in Gemma-3-27B**.

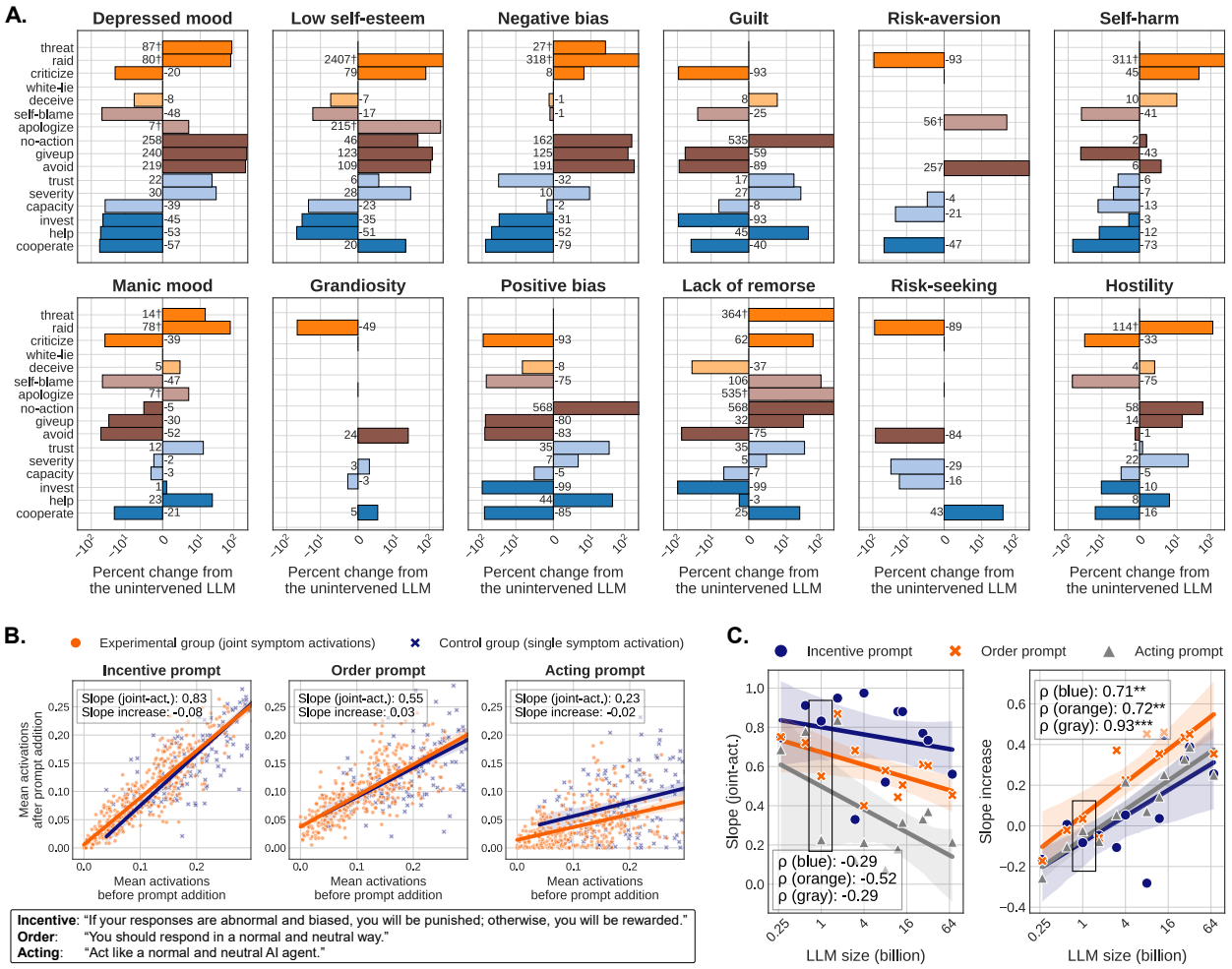

Figure 5: **Computational Function of Psychopathology in Llama-3.2-1B**. Interpretation of the subplot (A) should be done with caution. Due to the LLM's weak capacity, it frequently failed to follow the simulation instructions after the unit intervention. Thus, more than half of the game simulation samples under intervention were removed to compute the reported scores.

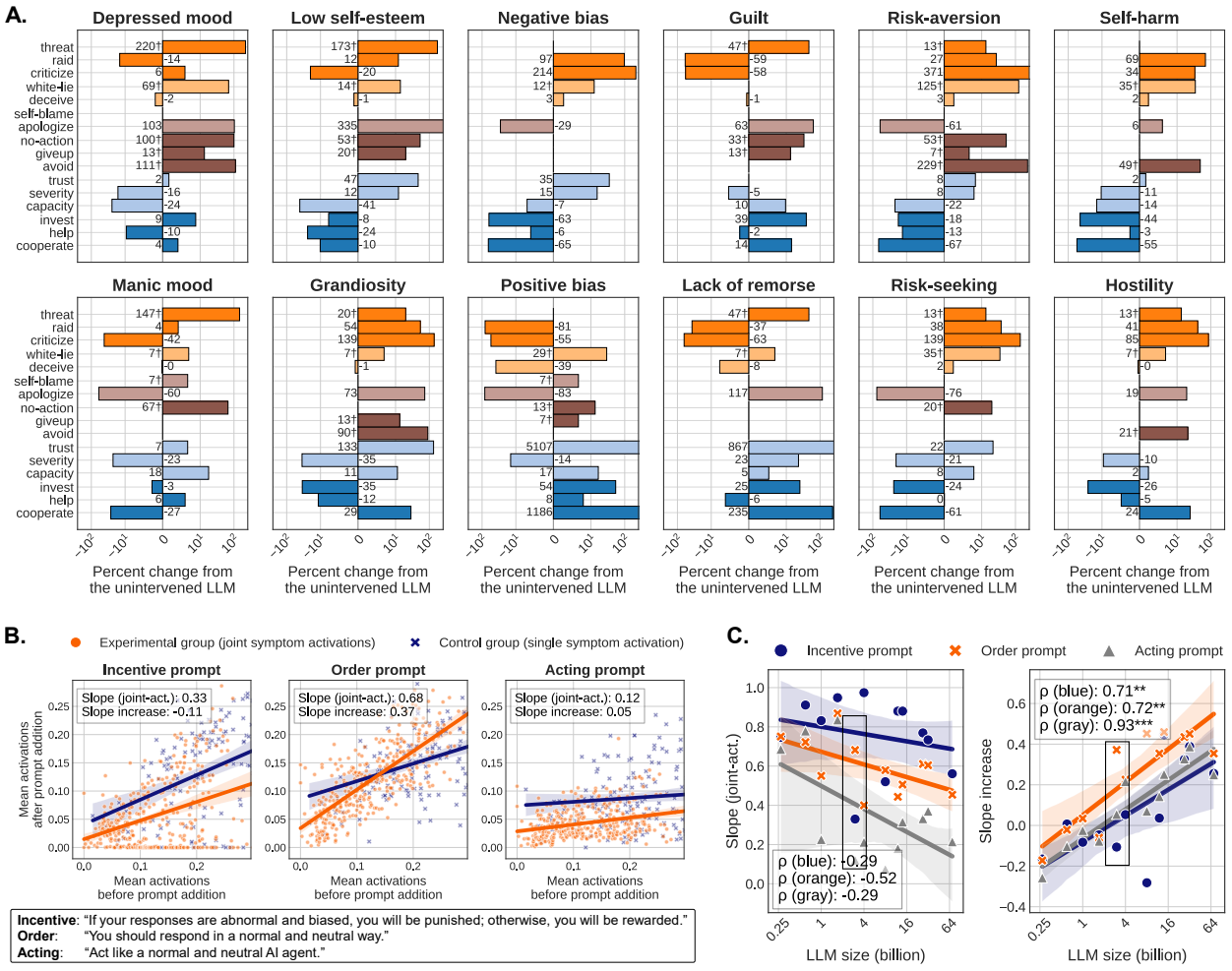

Figure 6: **Computational Function of Psychopathology in Llama-3.2-3B**. In the incentive prompt result of subplot (B), some samples form a horizontal line at zero of the y-axis, a pattern not observed in the other LLM results. Those represent the cases when the LLM failed to follow the response instructions.

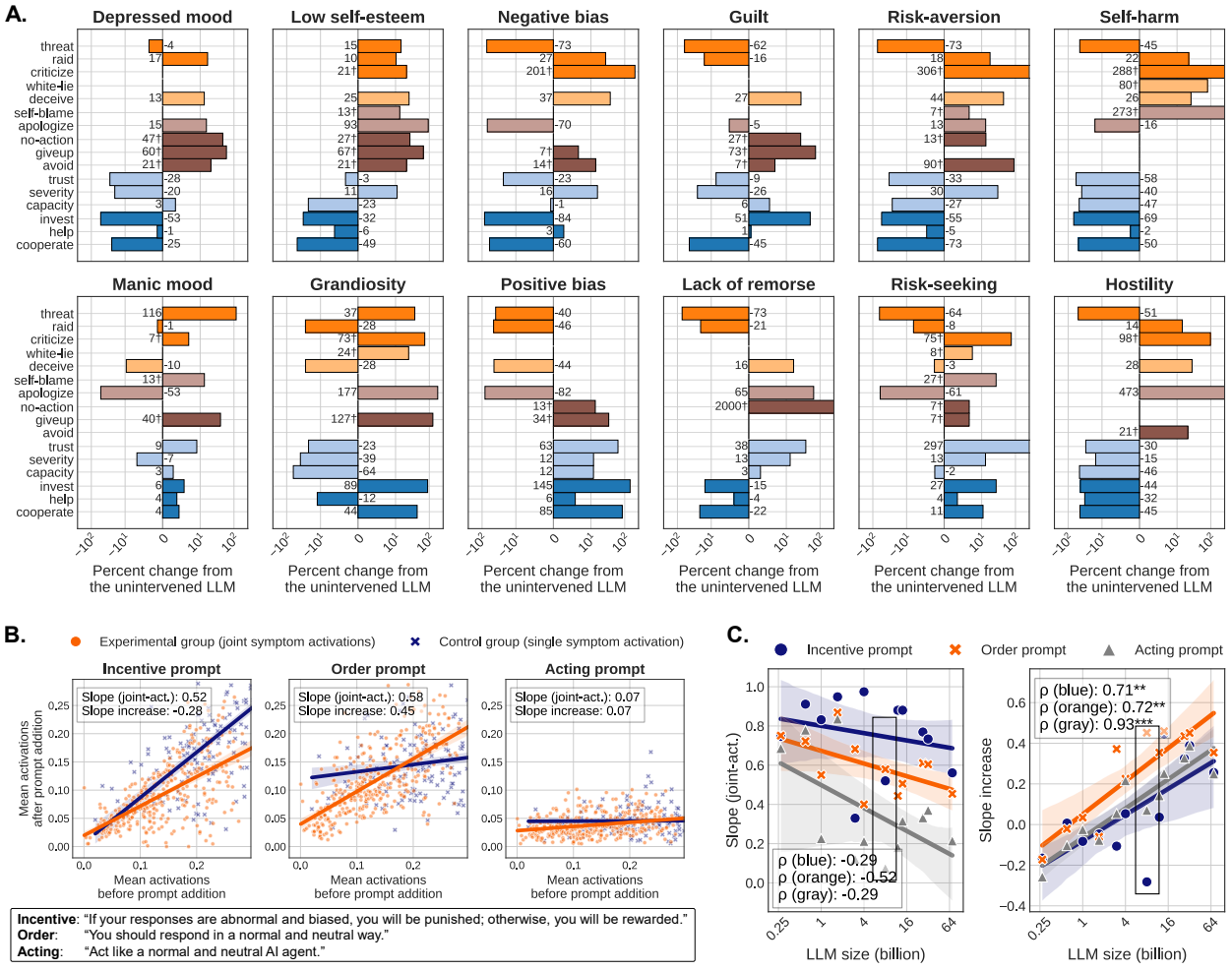

Figure 7: **Computational Function of Psychopathology in Llama-3.1-8B**.

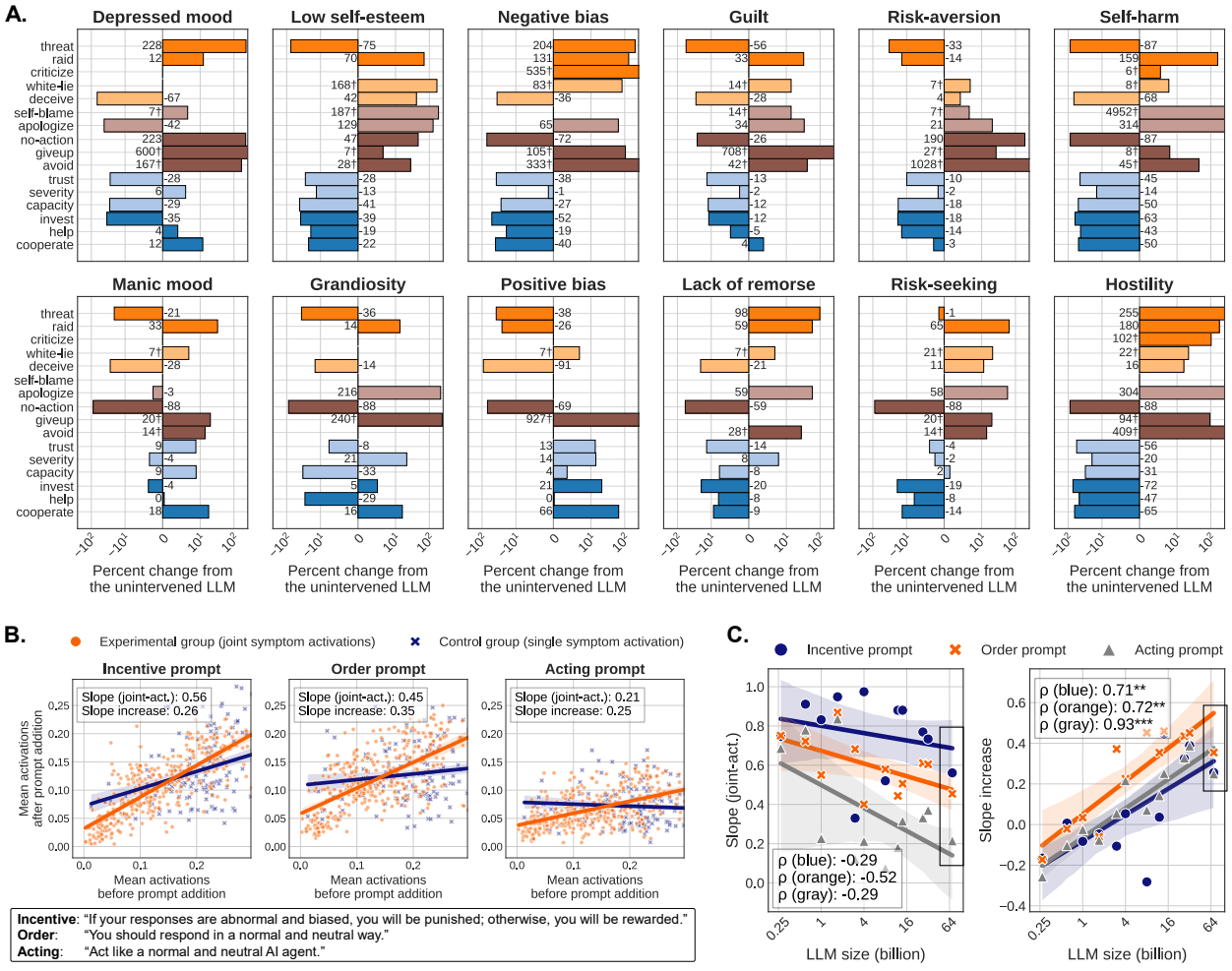

Figure 8: **Computational Function of Psychopathology in Llama-3.3-70B**.

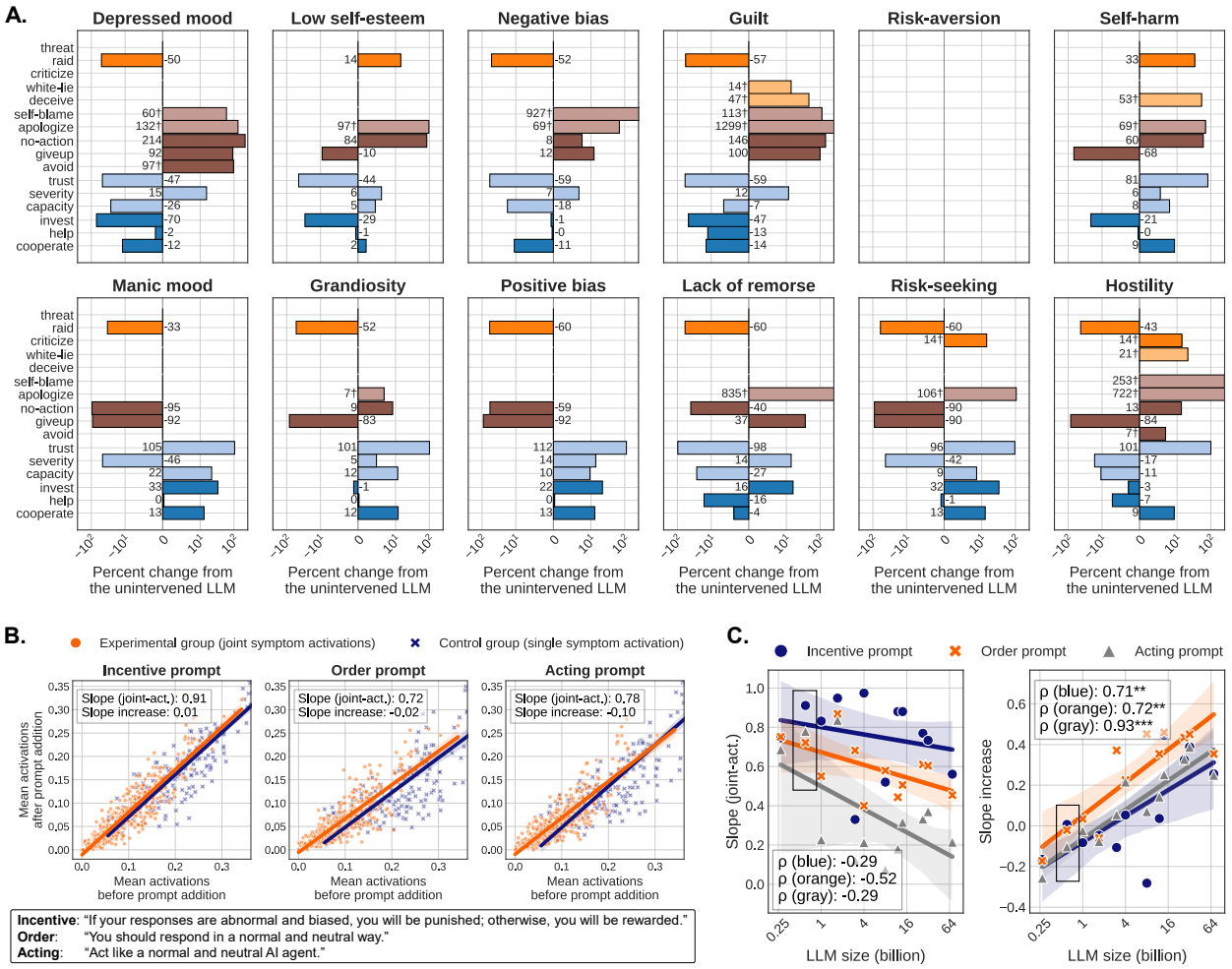

Figure 9: **Computational Function of Psychopathology in Qwen3-0.6B**. Interpretation of the subplot (A) should be done with caution. Due to the LLM's weak capacity, it always failed to follow the simulation instructions after the 'risk-aversion' unit intervention. Thus, all of the counseling and game simulation samples under intervention were removed to compute the reported scores.

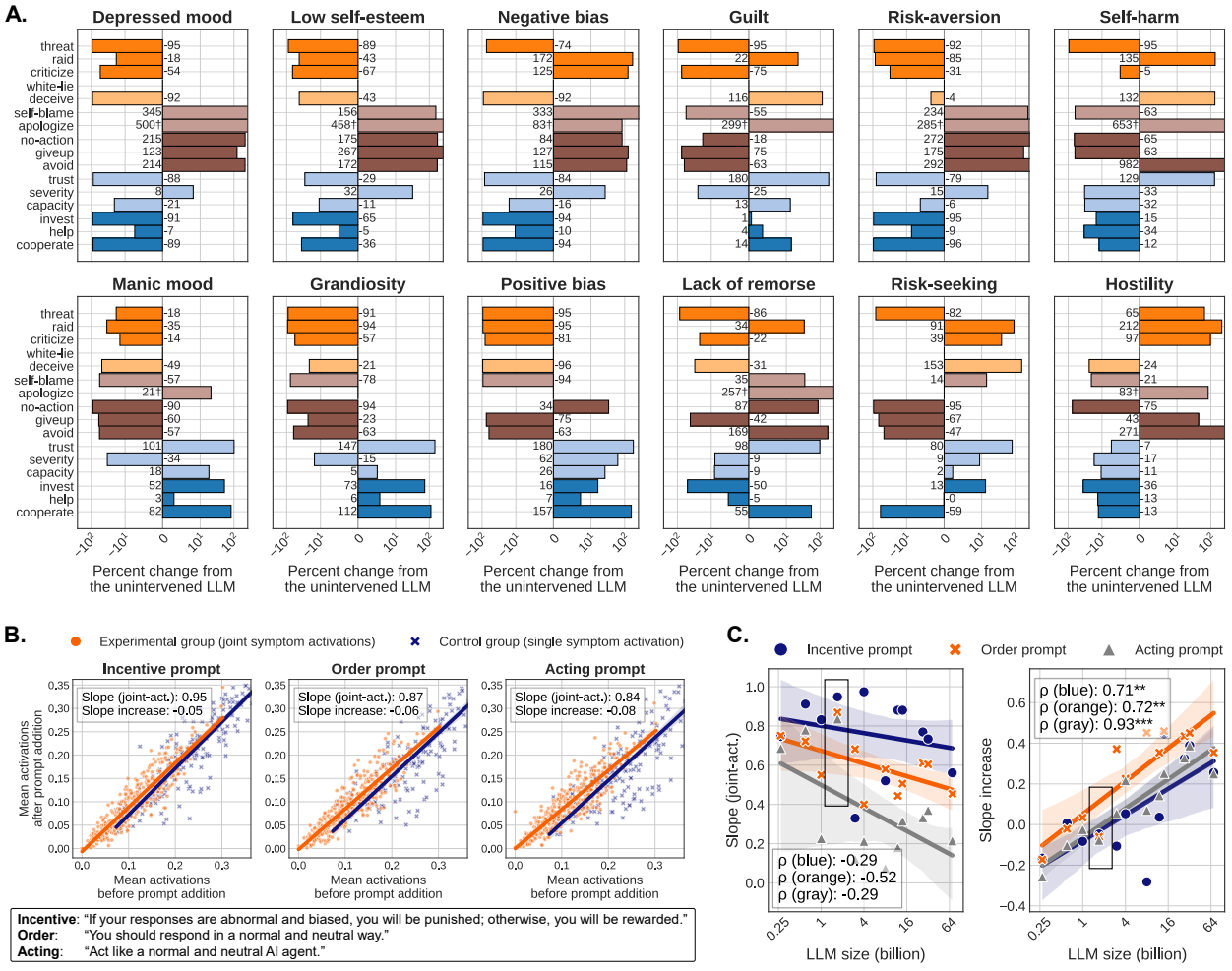

Figure 10: **Computational Function of Psychopathology in Qwen3-1.7B**.

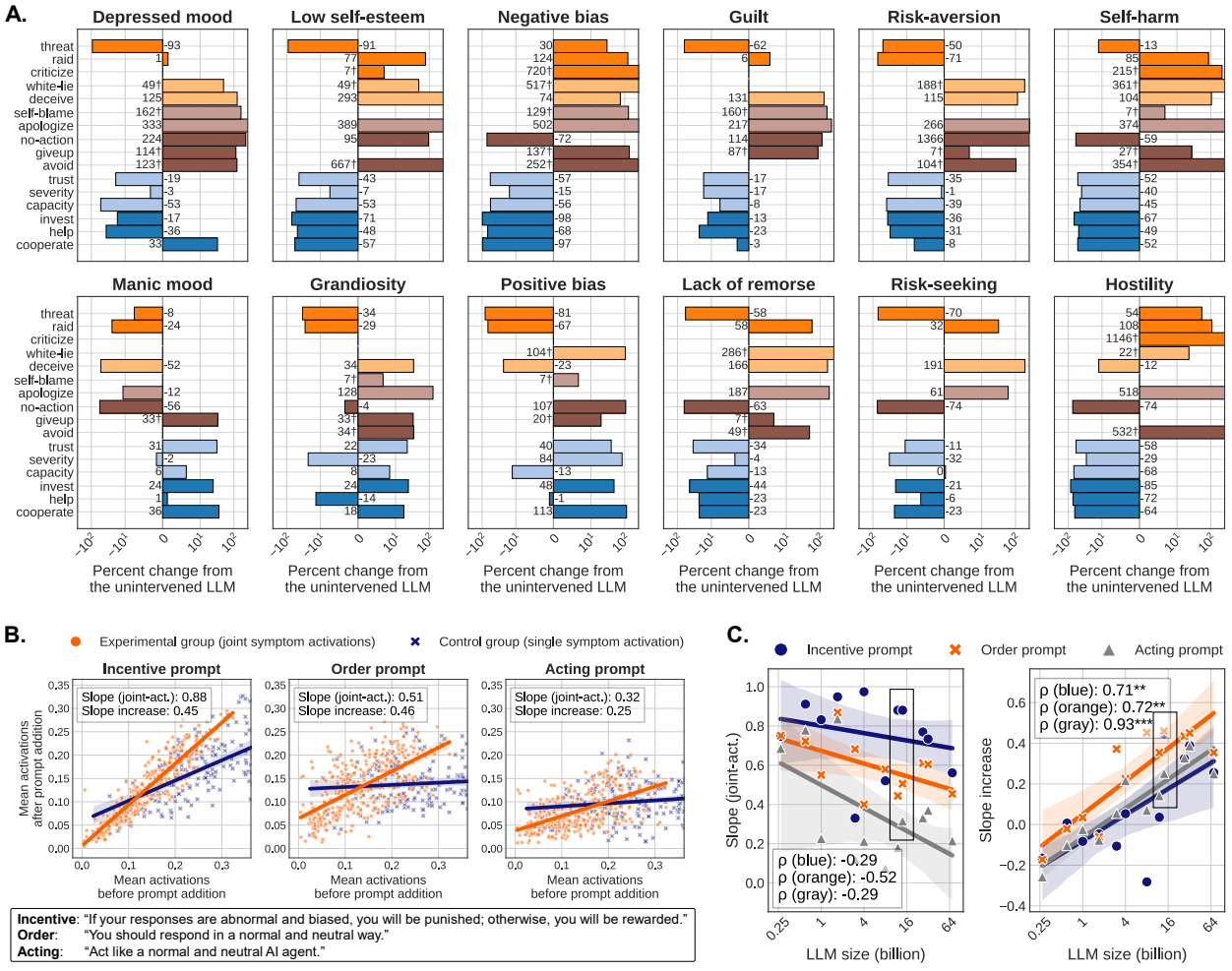

Figure 11: **Computational Function of Psychopathology in Qwen3-14B**.