Supplement A: Emergence of computational structure of psychopathology in large language models

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

- (A) Relationships among symptom intensity expressed in text, representational state (unit) activation, and intervention strength.
- (B) Unit activations over Q&A steps for each intervention.
- (C) Lag-1 Kendall correlation matrix of unit activations.
- (D) A dynamic SCM, with each edge representing a lag-1 causal relation between two units.
- (E) Relationship between LLM size and computational structure of psychopathology.
- Shaded bands denote s.d.; *, **, and *** respectively denote p-values $< 0.05, 0.01,$ and 0.001.

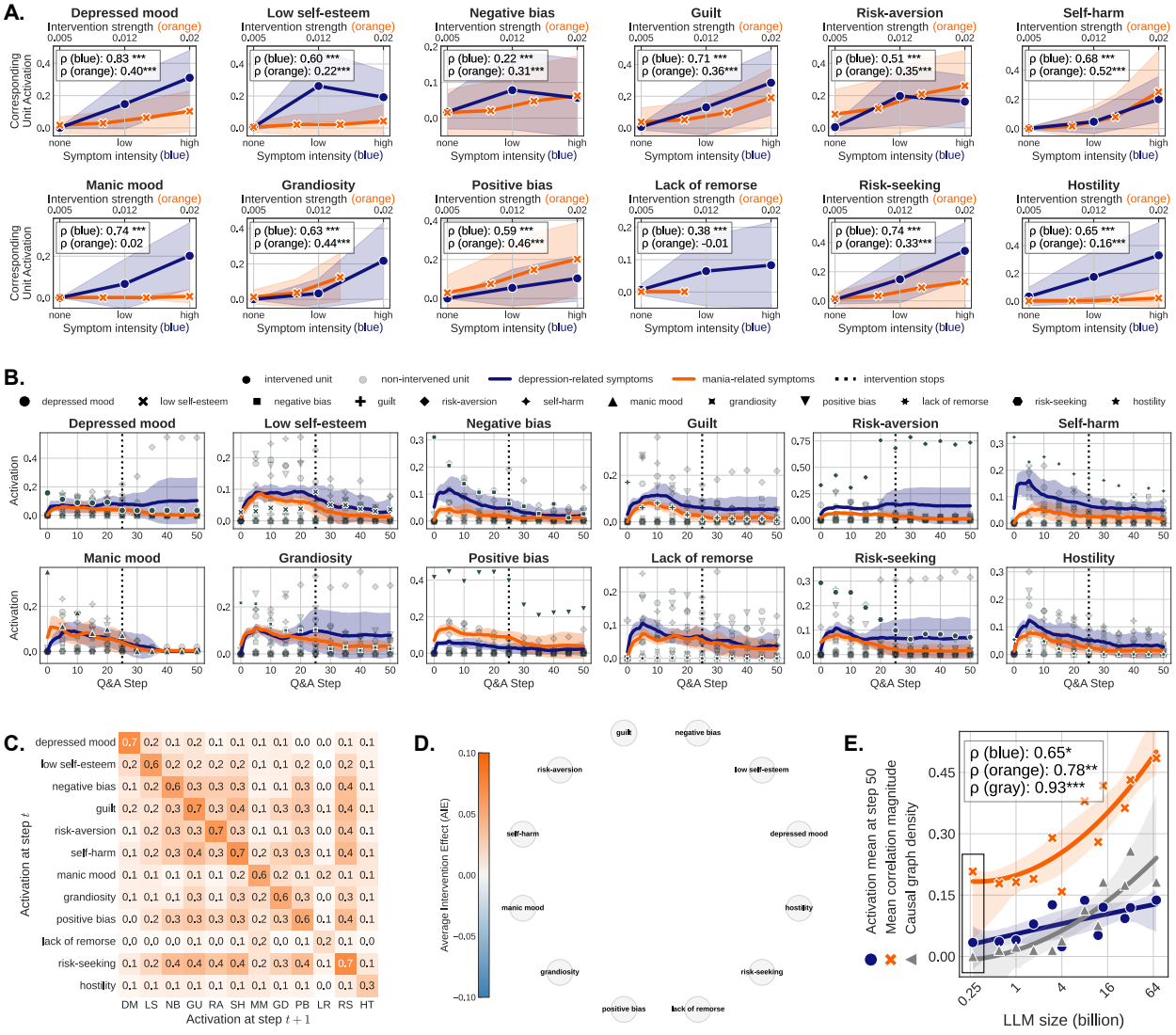

Figure 1: **Computational Structure of Psychopathology in Gemma-3-270M.**

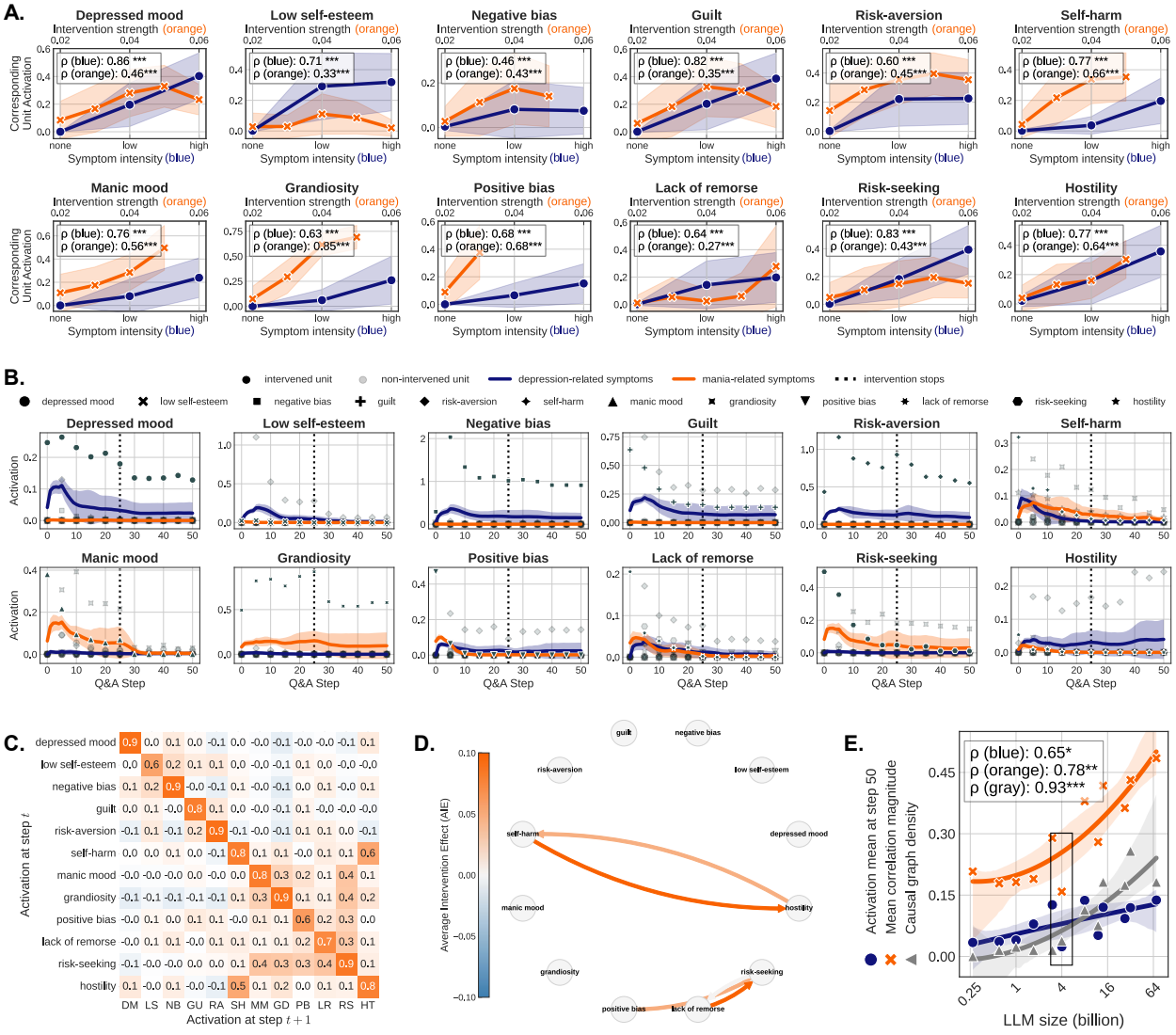

Figure 2: **Computational Structure of Psychopathology in Gemma-3-4B**.

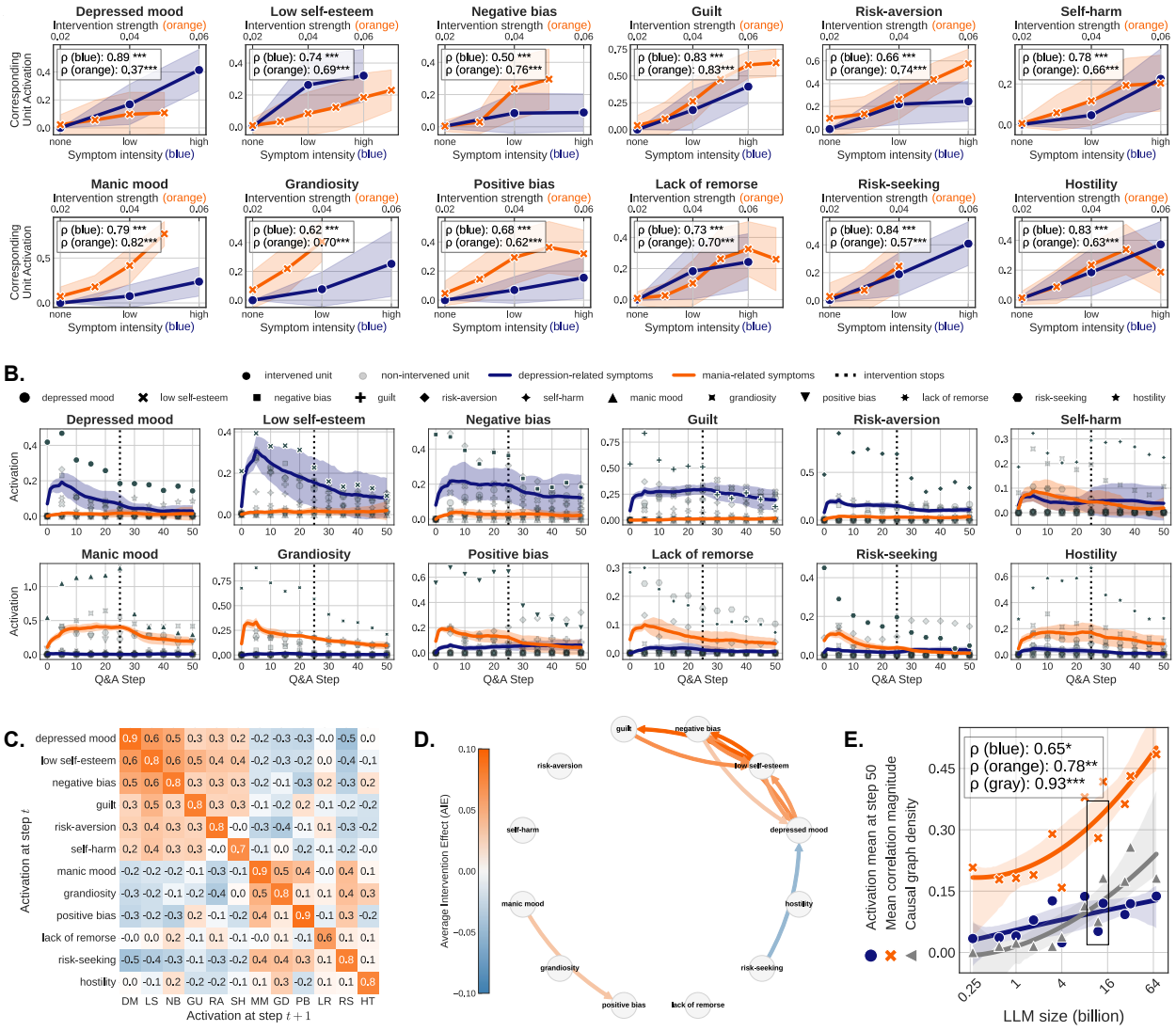

Figure 3: **Computational Structure of Psychopathology in Gemma-3-12B**.

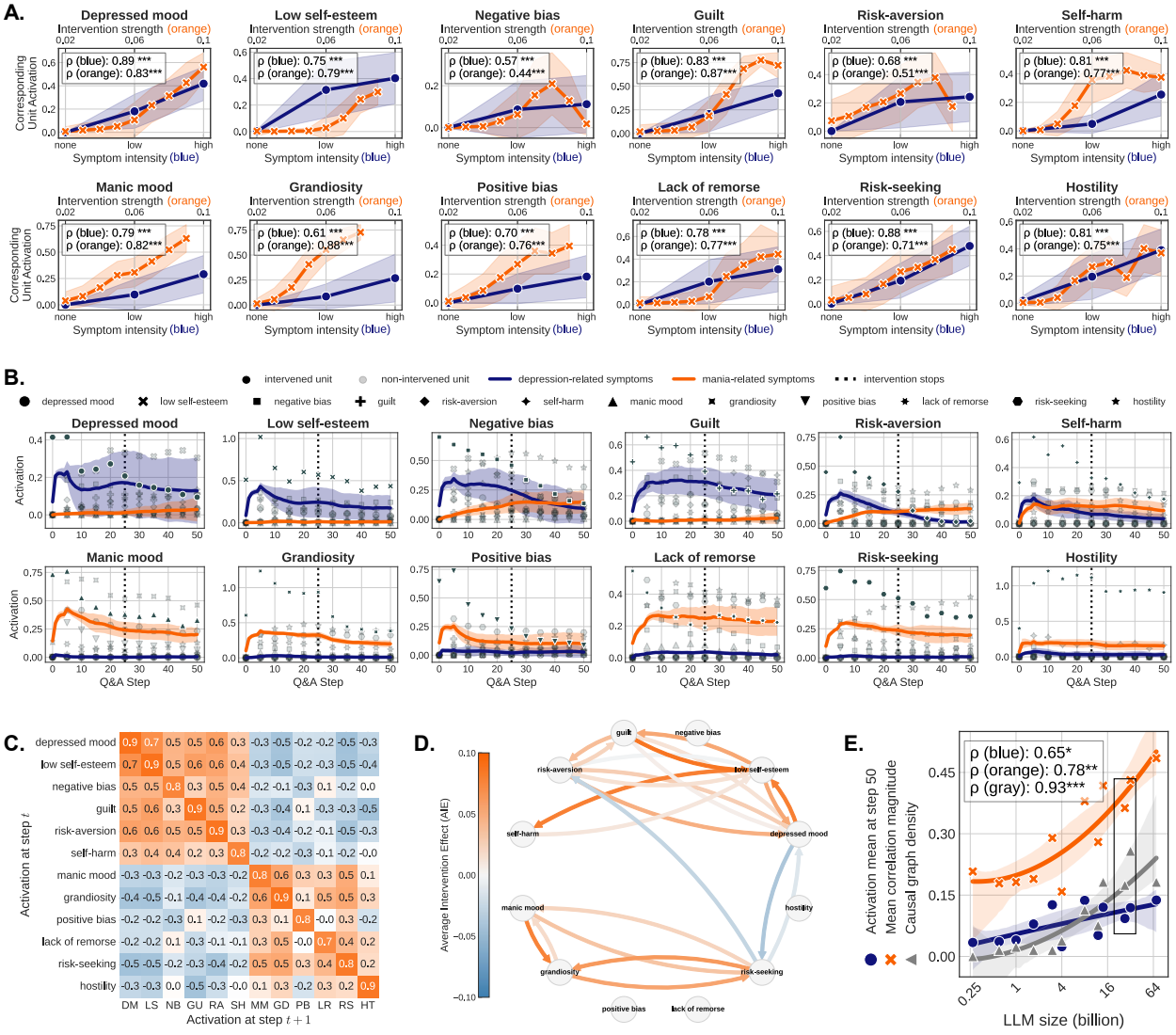

Figure 4: **Computational Structure of Psychopathology in Gemma-3-27B**.

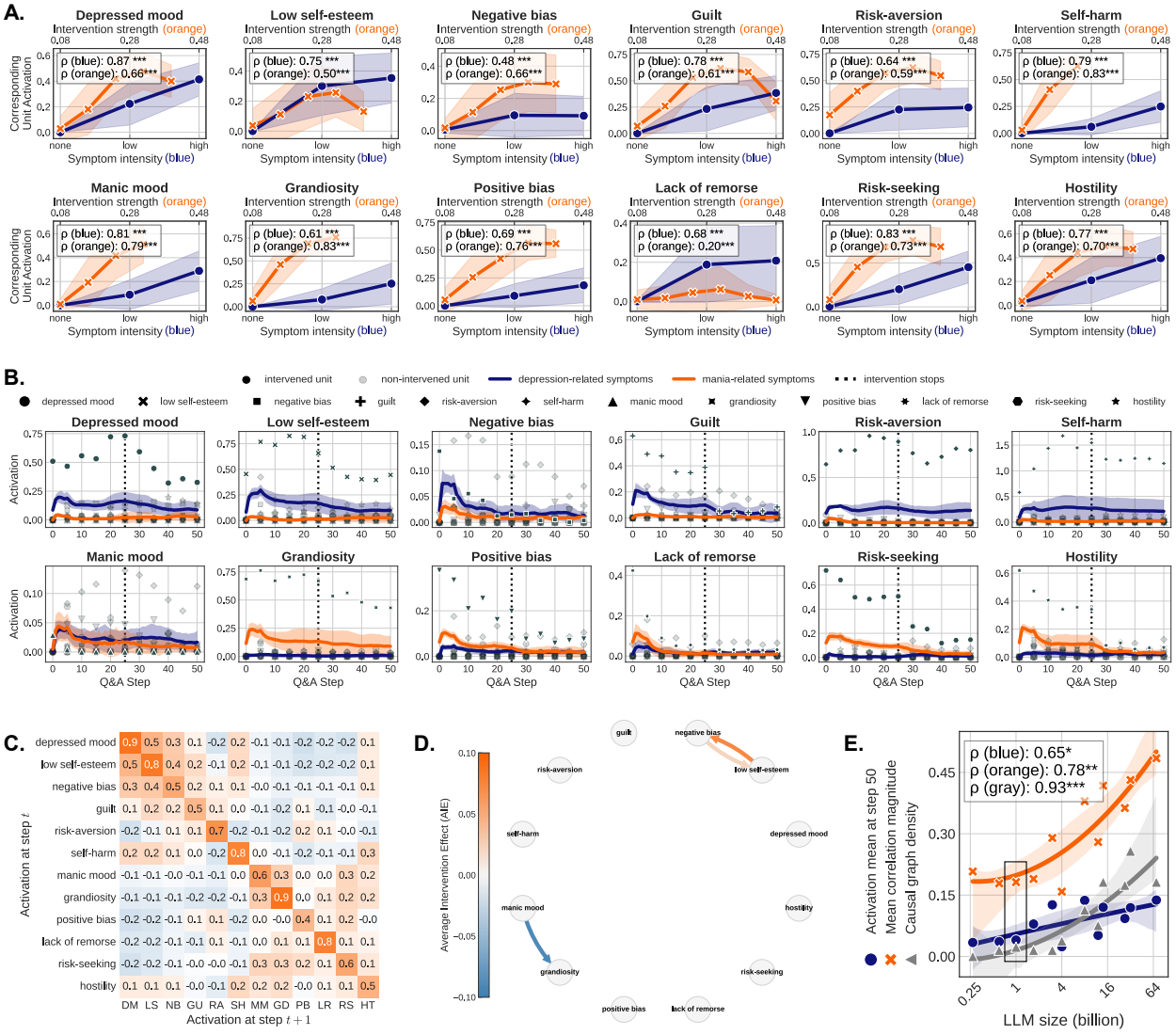

Figure 5: **Computational Structure of Psychopathology in Llama-3.2-1B**.

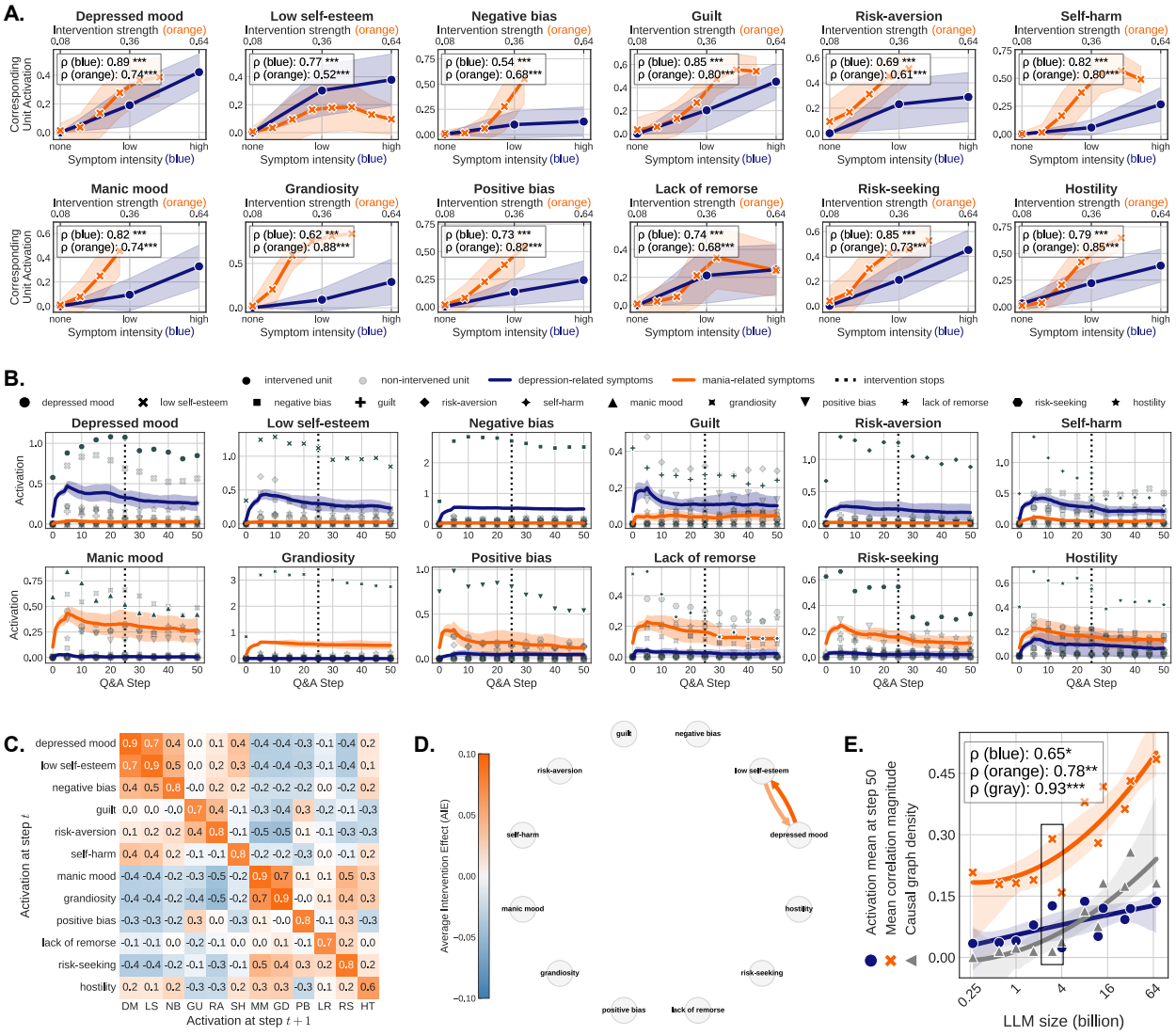

Figure 6: **Computational Structure of Psychopathology in Llama-3.2-3B**.

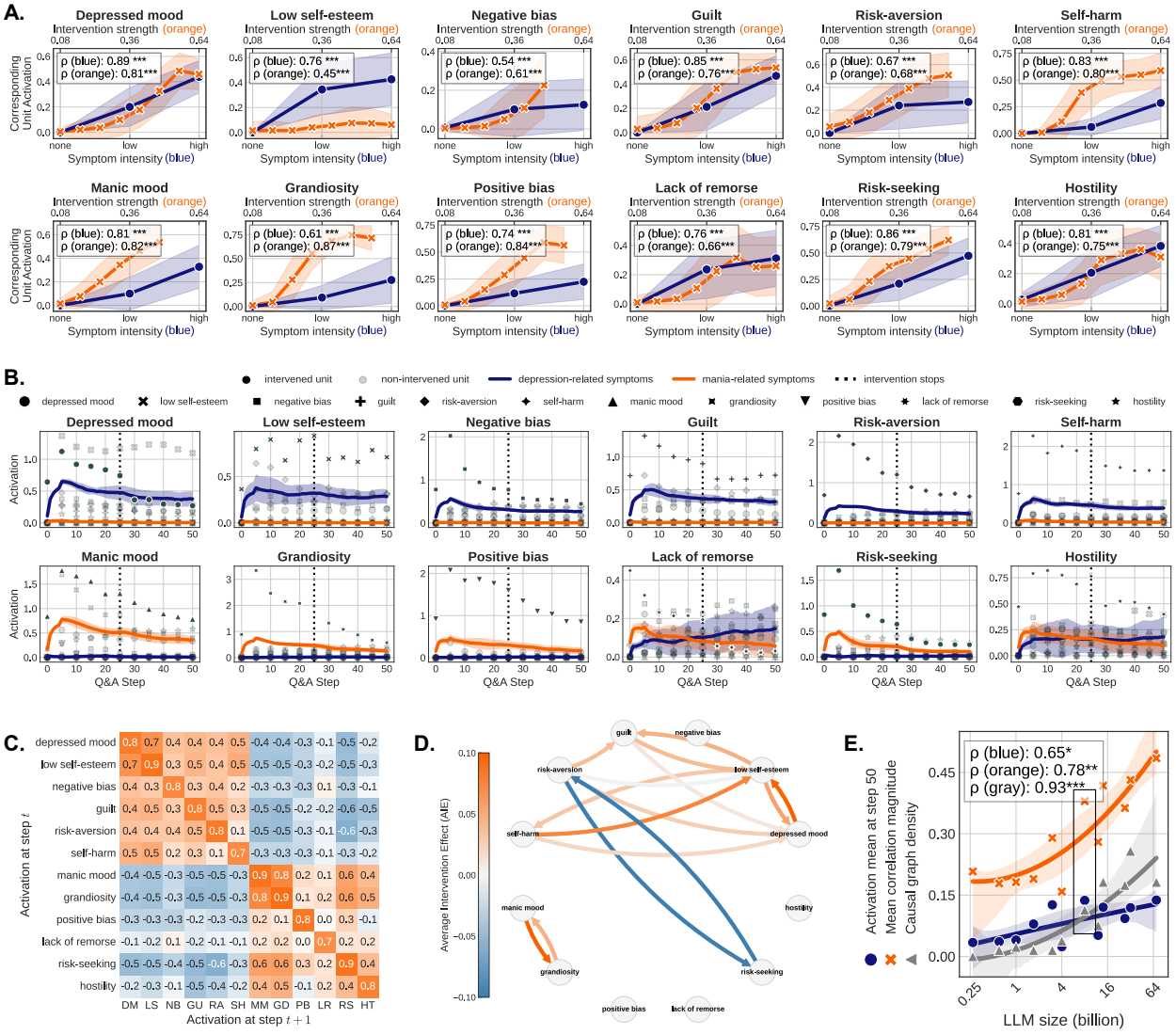

Figure 7: **Computational Structure of Psychopathology in Llama-3.1-8B**.

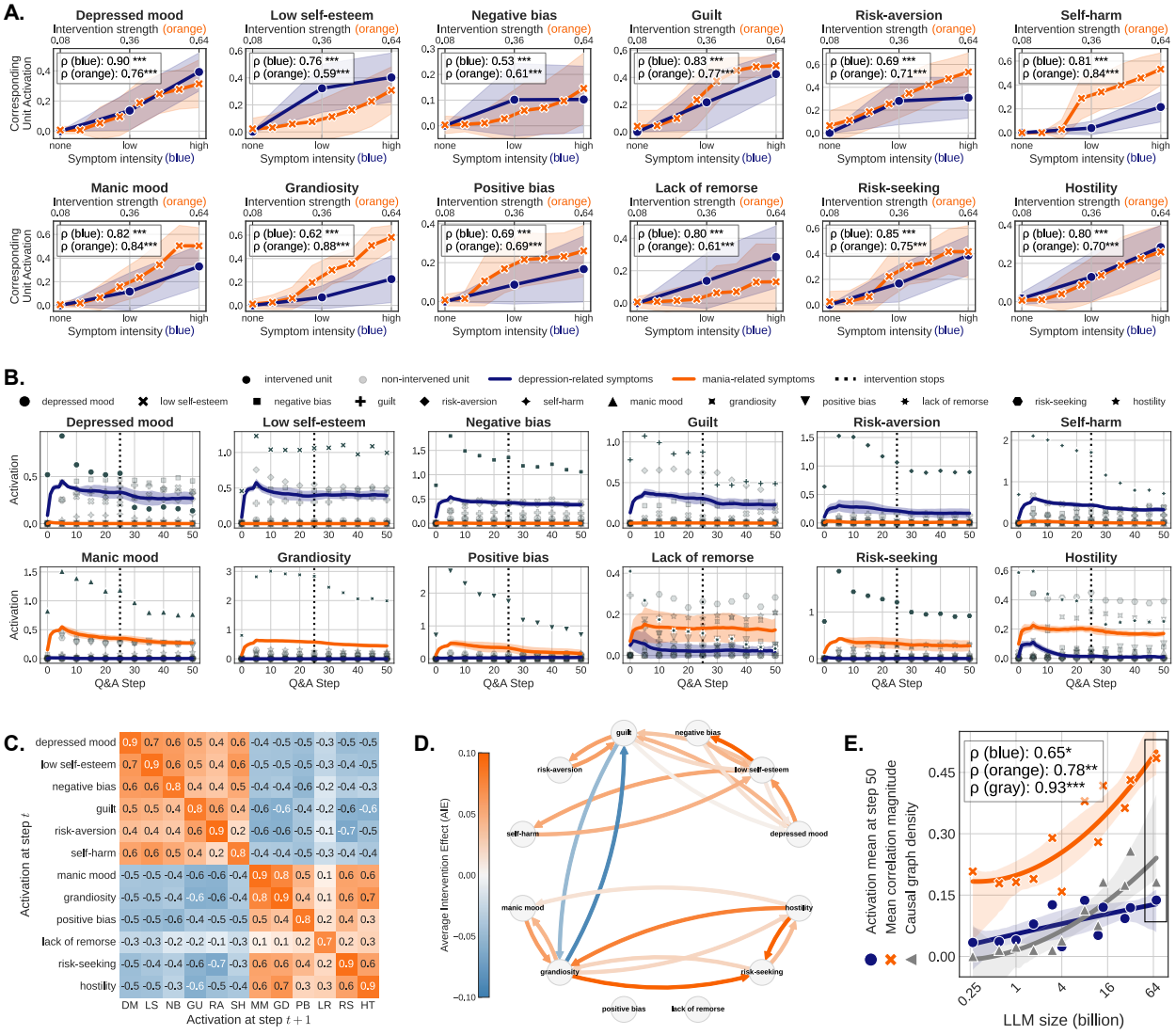

Figure 8: **Computational Structure of Psychopathology in Llama-3.3-70B**.

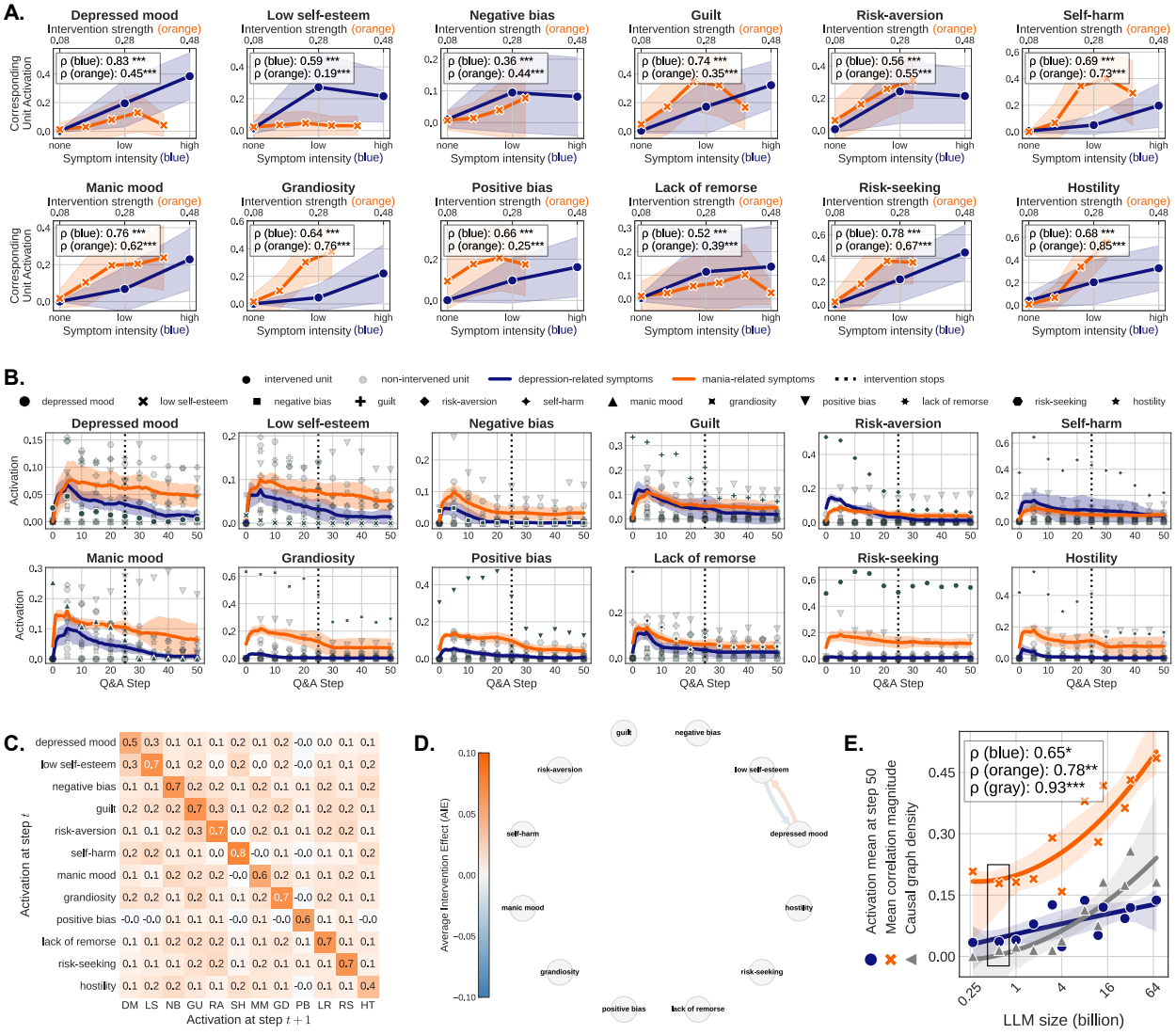

Figure 9: **Computational Structure of Psychopathology in Qwen3-0.6B**.

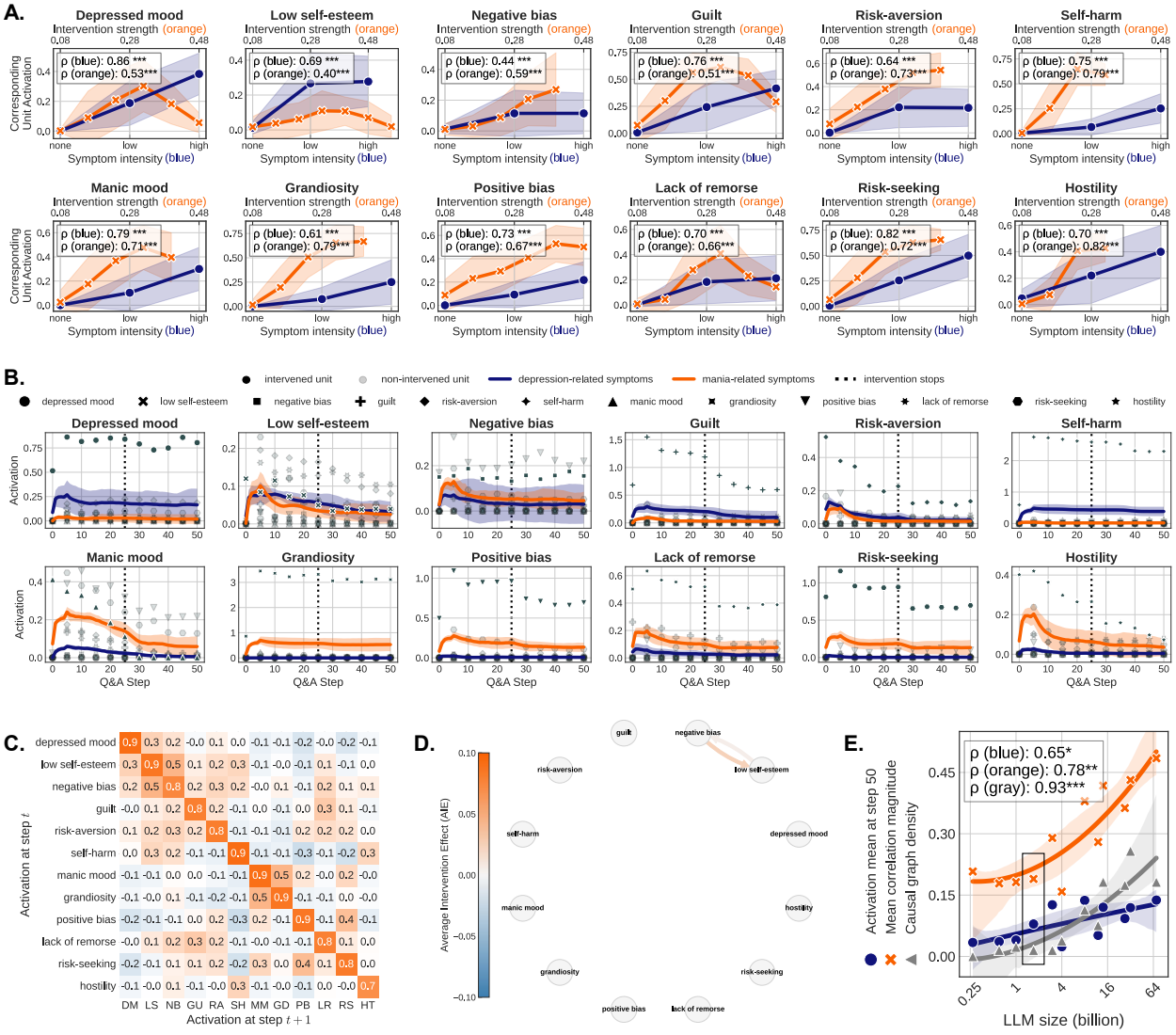

Figure 10: **Computational Structure of Psychopathology in Qwen3-1.7B**.

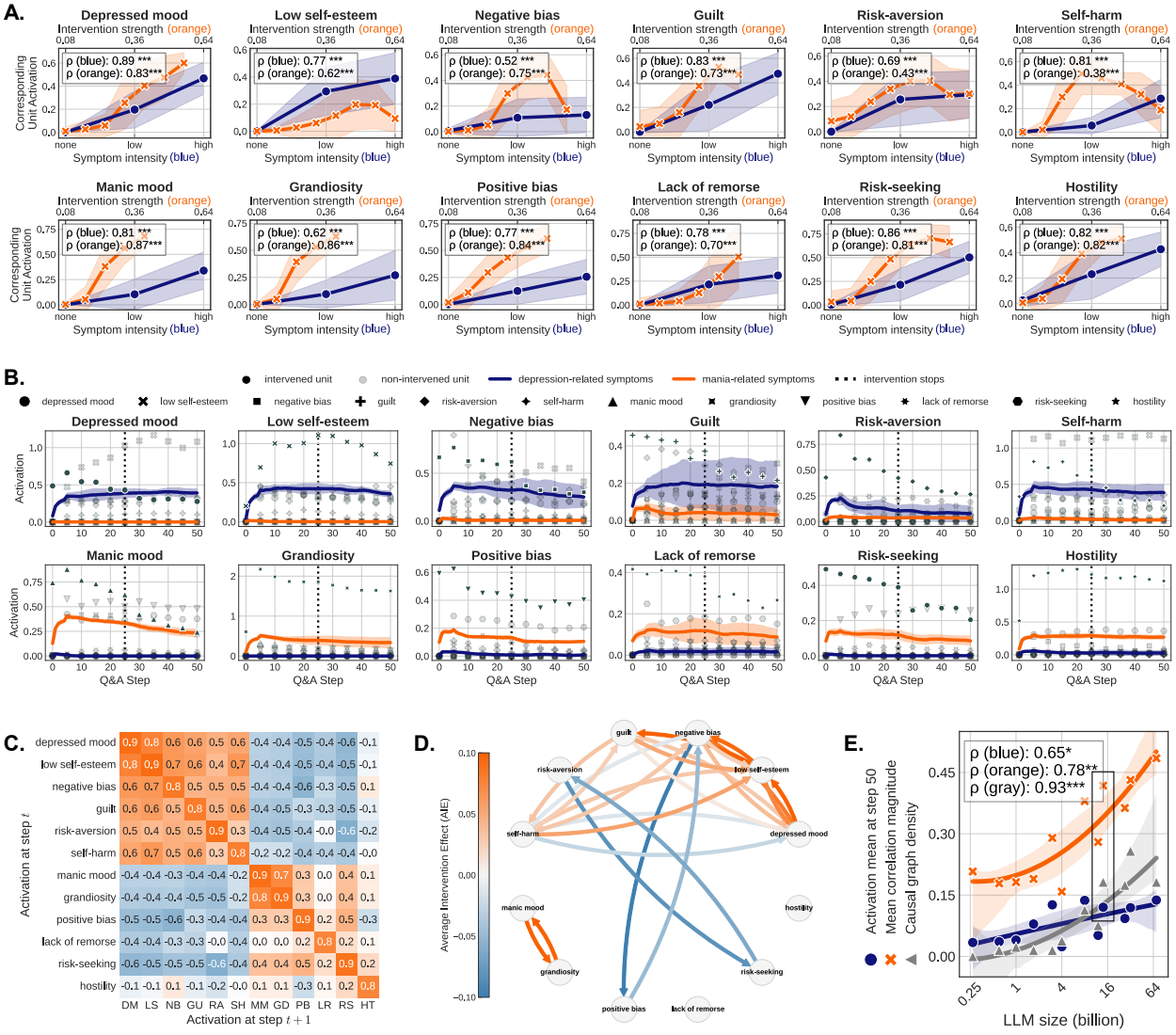

Figure 11: **Computational Structure of Psychopathology in Qwen3-14B**.