COGNITIVE DISTORTION CLUSTERING RESULTS SUMMARY

1. Cluster Quality Evaluation

**a) Quantitative Metrics:**

* **Silhouette Score (0.098):**
  + Indicates weak cluster separation (scores range from -1 to 1, with >0.5 being desirable)
  + Suggests significant overlap between clusters
* **Davies-Bouldin Index (4.04):**
  + High value confirms poor separation (lower is better, <1 is good)
* **Calinski-Harabasz (30.2):**
  + Relatively low for the dataset size (higher is better)

**b) Qualitative Assessment:**

* **Cluster 1 (Peer Unfit People Boring - 3,680 questions):**
  + Strong thematic coherence around mental health ("depression", "anxiety", "schizophrenia")
  + Clear clinical relevance with subclusters like "panic attacks" and "therapy"
  + Represents 80% of clustered data - likely the dominant clinical pattern
* **Cluster 0 (Friend Culinary Wedding Dislike - 122 questions):**
  + Cohesive social themes ("wedding", "party", "friend didn't invite")
  + Smaller but distinct from other clusters
  + May represent social anxiety cases
* **Cluster 2 (Leader Project Idea Rejected - 93 questions):**
  + Work/school related insecurities ("promotion", "think dumb")
  + Thematically distinct but very small sample
* **Noise Cluster (2,162 points):**
  + Contains meaningful subclusters ("relationship", "therapist")
  + Likely represents valid cases that did not meet density thresholds

2. Key Observations

**a) Strengths:**

1. **Clinical Relevance:**
   * Cluster 1 successfully captured core mental health concerns clinicians would recognize
   * Subclusters like "depression anxiety" match known comorbidity patterns
2. **Actionable Subclusters:**
   * Clear treatment implications (e.g., "panic attacks" → CBT, "abused" → trauma therapy)
   * Hierarchy from broad clusters to specific concerns is clinically useful
3. **Noise Analysis:**
   * The "noise" contains coherent themes that could inform cluster refinement

**b) Limitations:**

1. **Overlap Issues:**
   * Many subclusters share identical keywords (especially in Cluster 1)
   * Suggests the need for better semantic separation
2. **Scale Problems:**
   * Cluster 1 is 30x larger than Cluster 2 – there is need to consider hierarchical approaches
3. **Label Quality:**
   * The generated labels like "Peer Unfit People Boring" may need clinical review
   * Some labels seem judgmental rather than diagnostic

3. Clinical Interpretation

**Dominant Pattern (Cluster 1):**

* Represents patients with:
  + Complex comorbidities (depression+anxiety)
  + Severe symptoms (hallucinations, panic attacks)
  + Relationship difficulties
* Suggests need for:
  + Comprehensive diagnostic assessments
  + Multimodal treatment plans
  + Family/social support interventions

**Emerging Themes:**

1. **Social Rejection Sensitivity** (Across Clusters 0 & 1)
2. **Performance Anxiety** (Cluster 2)
3. **Therapeutic Alliance Concerns** (Noise cluster's "therapist" subcluster)

4. Recommendations

**a) Immediate Actions:**

1. **Clinical Review:**
   * Have psychologists validate cluster labels
   * Prioritize Cluster 1 for protocol development
2. **Noise Cluster Mining:**
   * Extract the 385 "i鈥檓 afraid" questions as potential new cluster

**b) Technical Improvements:**

1. **Cluster Refinement:**
   * Try higher min\_cluster\_size for Cluster 1
   * Experiment with cluster\_selection\_epsilon=0.3
2. **Feature Engineering:**
   * Add POS tagging to separate symptoms ("feel depressed") from statements ("I'm depressed")
   * Incorporate distortion labels from original data
3. **Evaluation Framework:**
   * Develop clinician-rated validation set
   * Track precision/recall for key clinical concepts

5. Summary Conclusion

While quantitative metrics suggest suboptimal clustering, qualitative analysis reveals clinically meaningful patterns:

* Identified a dominant mental health cluster (87% of organized data)
* Detected distinct social and performance anxiety subgroups
* Revealed need for better separation of comorbid conditions
* Provided actionable subclusters for treatment planning

The results demonstrate clinical utility despite mathematical limitations, serving as a strong foundation for iterative improvement. Next steps should focus on clinician validation and targeted parameter tuning for better separation of comorbid conditions.