**COGNITIVE DISTORTION CLUSTERING RESULTS: UPDATED SUMMARY**

**1. Cluster Quality Evaluation**

**a) Quantitative Metrics**

* **Silhouette Score (0.098)**: Weak separation (ideal >0.5)
* **Davies-Bouldin (4.04)**: Poor clustering (ideal <1)
* **Calinski-Harabasz (30.2)**: Low for dataset size

**b) Qualitative Assessment** *(Enhanced with Distortion Analysis)*:

* **Cluster 1 (Peer Unfit People Boring)**:
  + **Clinical Relevance**: Strong themes of depression, anxiety, psychosis.
  + **Distortion Profile**: 64.9% distorted (moderate), weak correlation (−0.064). Suggests *mixed severity*—some adaptive responses alongside maladaptive thoughts (e.g., "I’m scared" vs. "I cope sometimes").
  + Cluster has 20 subclusters, with a total of 3680 questions and an average of 184 questions per subcluster. Minimum questions is 53 maximum is 504.
* **Cluster 0 (Friend Culinary Wedding Dislike)**:
  + **Social Rejection Theme**: Small but coherent (e.g., "wasn’t invited").
  + **Distortion Profile**: 98.4% distorted (near-universal). Likely reflects *overgeneralization* ("Nobody likes me") or *mind-reading* ("They excluded me on purpose").
  + Has a total of 18 subclusters from 122 questions with an average of 7 questions, minimum of 1 and maximum of 17 questions per subcluster.
* **Cluster 2 (Leader Project Idea Rejected)**:
  + **Performance Anxiety**: Work/school insecurities ("think dumb").
  + **Distortion Profile**: 100% distorted. Indicates *rigid, extreme distortions* (e.g., "I’m incapable" → all-or-nothing thinking).
  + Has 19 subclusters from a total of 93 questions, average of 5 questions per subcluster with minimum of 1 and maximum of 21 questions.
* **Noise Cluster**:
  + **Mixed but Meaningful**: Subclusters like "therapist" (202 Qs) and "i’m afraid" (385 Qs).
  + **Distortion Profile**: 74.1% distorted, strongest correlation (−0.30). Suggests *predictable linguistic patterns* in distortions (e.g., repetitive catastrophizing).
  + Has 20 subclusters, total of 2162 questions, average questions per subcluster is 108, with a minimum of 14 and maximum of 385 questions

**2. Key Observations**

**a) Strengths** *(Augmented)*:

* **Clinical Utility**:
  + Cluster 1’s subclusters (e.g., "panic attacks") align with CBT targets.
  + Noise cluster’s strong distortion correlation (−0.30) helps flag *high-risk linguistic markers*.
* **Distortion Patterns**:
  + Universal distortion in Cluster 2 signals need for *urgent intervention* (e.g., cognitive restructuring).
  + Cluster 0’s near-total distortion highlights *social anxiety with pervasive negative bias*.

**b) Limitations** *(Updated)*:

* **Overlap Issues**: Shared keywords (e.g., "anxiety" in Cluster 1/noise) may reflect *comorbidity* rather than poor clustering.
* **Distortion Correlation Gaps**: Weak correlations in Clusters 0/1 suggest distortions are *less linguistically predictable*—may require clinician review.

**3. Clinical Interpretation** *(Enhanced)*

**Dominant Pattern (Cluster 1)**:

* **Profile**: Comorbid depression/anxiety, relational distress.
* **Distortion Insight**: Moderate distortion rate (64.9%) implies *varied symptom severity*—ideal for *graded interventions*.

**High-Risk Groups**:

* **Cluster 2 (100% Distorted)**: Priority for *CBT* (challenge all-or-nothing thinking).
* **Noise Cluster’s "I’m Afraid" (385 Qs)**: Likely *catastrophizing*—target with behavioral experiments.

**Emerging Themes**:

* **Social Rejection Sensitivity**: Cluster 0’s distortions (98.4%) + Cluster 1’s "relationship" subcluster (376 Qs).
* **Therapeutic Alliance Concerns**: Noise cluster’s "therapist" subcluster (202 Qs) may reflect *distrust* or *misattribution*.

**4. Recommendations** *(Expanded)*

**a) Clinical Actions**:

* **Urgent Review of Cluster 2**: 100% distortion rate warrants *risk assessment* (suicidality?).
* **Noise Cluster Mining**: Extract "i’m afraid" (385 Qs) as a *new high-distortion cluster*.

**b) Technical Improvements**:

* **Feature Engineering**: Add *distortion-type tags* (e.g., "catastrophizing") to improve separation.
* **Correlation Analysis**: Investigate why noise cluster distortions correlate strongly (−0.30) with linguistic features.

**c) Validation**:

* **Clinician Coding**: Manually label 100 samples per cluster for *distortion types* (e.g., overgeneralization vs. personalization).

**5. Summary Conclusion**

The integration of distortion metrics reveals:

1. **High-Risk Clusters**: Cluster 2 (100% distorted) and noise subclusters (e.g., "i’m afraid") need prioritized care.
2. **Linguistic Patterns**: Noise cluster’s strong correlation (−0.30) offers a *proxy marker* for severe distortions.
3. **Clinical Potential**: Despite weak quantitative metrics, clusters map to *meaningful clinical profiles* (e.g., social rejection vs. performance anxiety).

**Appendix: Subcluster & Correlation Highlights**

*(Refer to subcluster\_analysis\_report.txt for full details)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Key Subclusters** | **Top Distortion Indicators** | **Pearson (r)** |
| **Cluster 1** | "i’m scared" (504 Qs) | Mixed severity (−0.064) | −0.064 |
| **Cluster 0** | "wasn’t invited" (12 Qs) | Near-universal (−0.047) | −0.047 |
| **Noise** | "i’m afraid" (385 Qs) | Strong linguistic link (−0.304) | **−0.304** |
| **Cluster 2** | "project didn’t" (21 Qs) | Universal distortion (NaN) | NaN |

**Insight**: The noise cluster’s stronger correlation suggests its distortions are more *formulaic* (e.g., repetitive phrases), while other clusters may reflect *diverse cognitive biases*.

This summary ties quantitative clustering results to *clinical interpretations of cognitive distortions*, highlighting actionable insights for therapy and research.