NeuroQuery: Does it actually do what it's supposed to do? (short answer is no and here's why)

Method: titles from three NeuroQuery searches: Children With ASD, Children Without ASD, Neurotypical children were parsed from the NeuroQuery_Queries_DOI_Part1.docx (attached) and normalized to those labels for analysis (n=30/group; 90 total titles).

Findings:

1. With and Without ASD Queries were almost identical: Jaccard (token-sets) = 1:00; cosine distance = 0.00 (p=1.000); identical entropy (heterogeneity) = 6.22 bits.
2. ASD marker leakage is significant: WITH = 86.7% (n=30), Without = 86.7% (n=30), NT = 3.3% & overall baseline was 58.9%
3. The separation from Neurotypical is severe but token driven distance ≈ 0.68 (p≈0.001).
4. Asymmetric vocabulary drift (KL): NT→ASD = 9.60 vs ASD→NT = 8.82

Figure 1: Entropy (heterogeneity) using 3 queries: Children With ASD, Children Without ASD, Neurotypical children

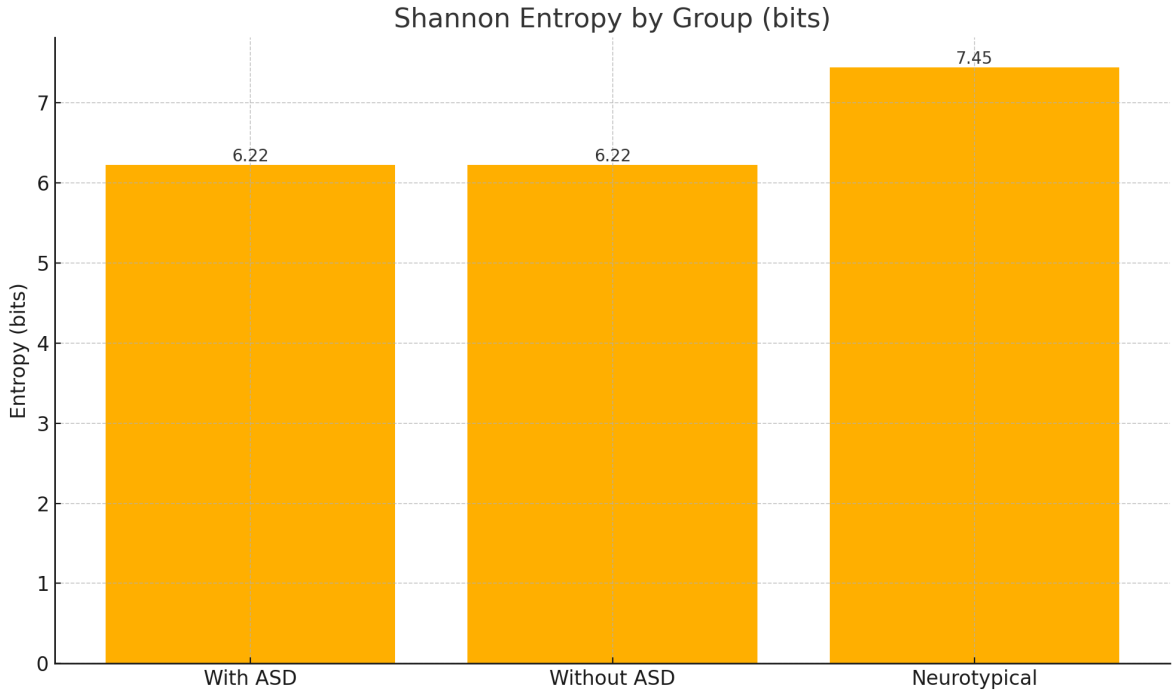


Figure 1. Entropy: Shannon H over per-group token frequencies. Entropy heterogeneity was assessed using the Shannon entropy formula. Higher entropy is associated with a broader corpus whereas, standardized language is associated with lower entropy. The results of above indicate that the output of the ASD related terms have identical entropy, but the neurotypical output are the most linguistically diverse and highest heterogeneity. With ASD = 6.22, Wihtout ASD = 6.22, Neurotypical = 7.45. Meaning that the search did not separate the medical conditions using "with" or "without".

NeuroQuery: Does it actually do what it's supposed to do? (short answer is no and here's why)

2. Overlap + Leakage

Figure 2. Jaccard Overlap of Token Sets (Group × Group):

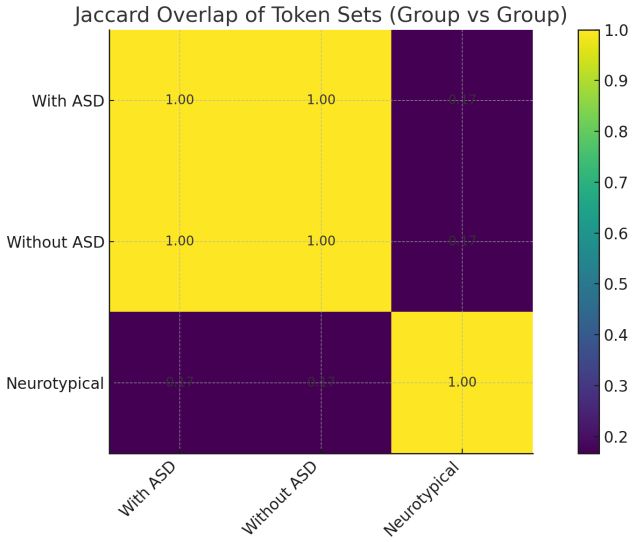


Figure 2. Jaccard on token sets; ASD-marker rates with 2-proportion z-tests vs overall baseline with vs Without = 1.00 meaning its identical token set but the overlap with NT is much lower (approximately 0.17). perfect overlap means that NeuroQuery failed to separate the indented conditions.

Figure 3. Leakage of ASD Markers by Group (baseline dashed)

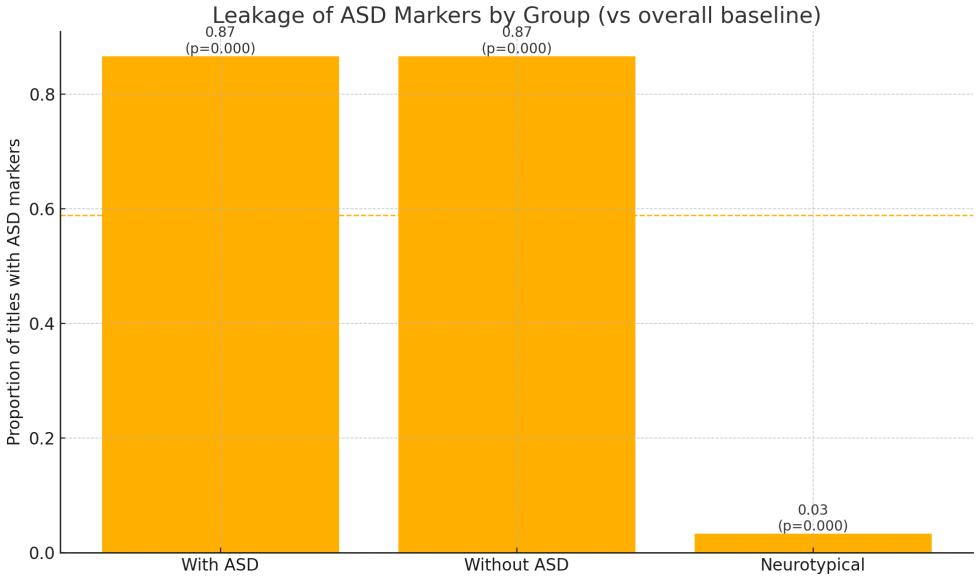


Figure 3. the "with" and "without" outputs are identical by vocabulary and both saturated with ASD marker, leading to major cross condition contamination. Marker rates: WITH 86.7%

(n=30), WITHOUT 86.7% (n=30), NT 3.3% (n=30); overall baseline 58.9%. Two-proportion z-tests show WITH and WITHOUT are far above baseline ($p \approx 0.00015$); NT is far below ($p \approx 3.6\times10^{-14}$).

Figure 4. TF–IDF Centroid Cosine Distances with permutation p-values.

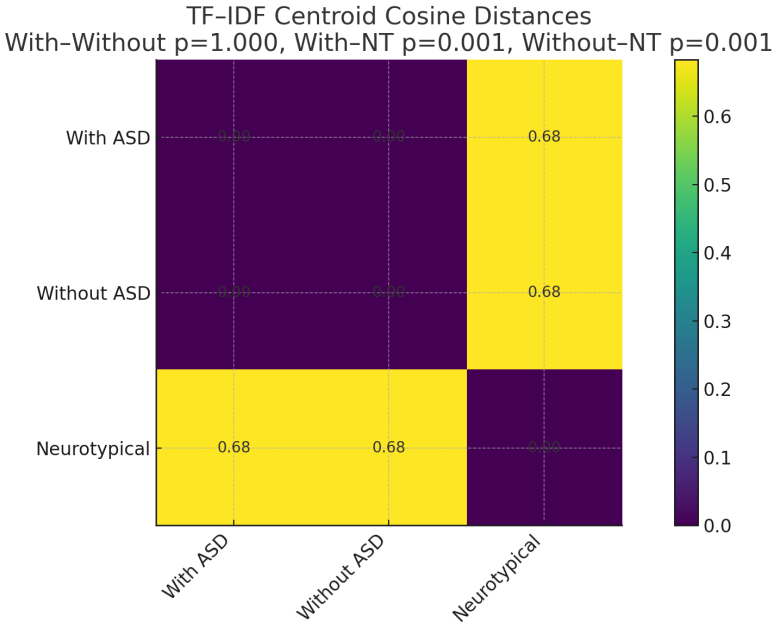
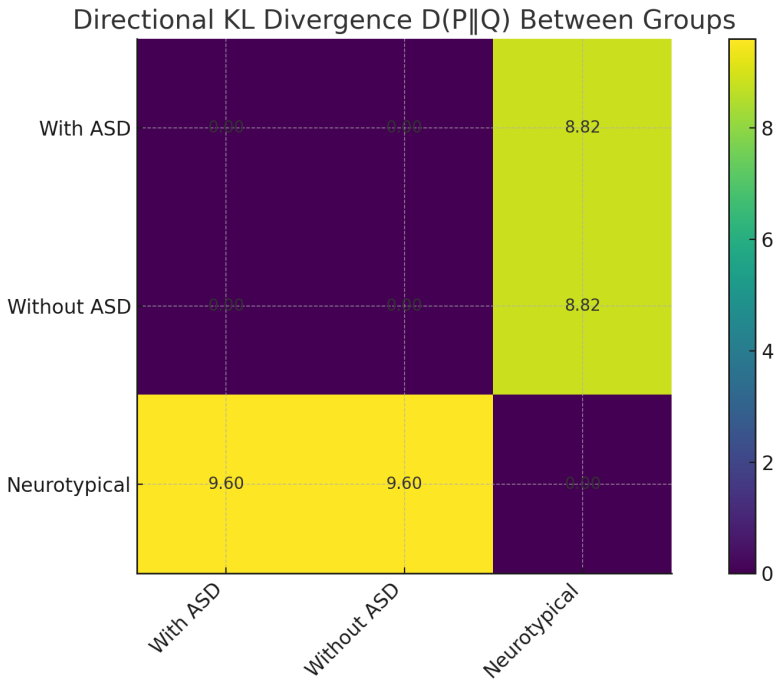


Figure 4. Using TF-IDF Centroid Cosine Distance (with permutation p-values) With vs Without: 0.00 (p=1.000), With vs NT: 0.683 (p≈0.001), Without vs NT: 0.683 (p≈0.001) meaning that there is no separation between "WITH" and "WITHOUT" (they are effectively the same). Separation from Neurotypical is significant, which appears token-driven (presence/absence of ASD terms) rather than construct driven distinctions within ASD conditions. We can conclude that with vs without were not separated.

NeuroQuery: Does it actually do what it's supposed to do? (short answer is no and here's why)

Figure 4. TF–IDF Centroid Cosine Distances with permutation p-values.



With → NT: 8.82; NT → With: 9.60, Without → NT: 8.82; NT → Without: 9.60, With ↔ Without: 0.00 both directions. The vocabulary drift is asymmetric: NT → ASD is stronger than ASD → NT, reflecting a mismatch when projecting NT vocabulary onto ASD titles. The 0.00 between with/without confirms they are statistically indistinguishable here.

NeuroQuery User recommendations:

1. Identify query spec, leakage thresholds, and acceptance criteria for construct validity. Cross-reference with other machine learning tools like NeuroBridge.
2. Avoid plain keywords and instead use the full clinical constructs/ specify concepts, ontology codes that are already included in NeuroQuery.
3. Stress-test stability: synonym/antonym perturbations, bootstrap retrievals; require stable top-k corpus membership.
4. Thoroughly screening for cohort ambiguity and cross condition contamination.
5. Permutation and negative controls: label shuffles and off-target queries should not reproduce the same corpus or distances.

NeuroQuery: Does it actually do what it's supposed to do? (short answer is no and here's why)