LLM을 활용한 유사 의미 군집 결과 데이터 검토

: 총 4개의 모델을 이용한 데이터 군집화 결과, 공통된 10개의 군집명을 선정하여 검토 진행.

: 선정한 군집명 10개는 아래 참고,

< Representative >

- 10% of admissions to regular hospitals due to non-compliance
- age-related disabilities and diseases
- infertility
- difficulty in transporting the lifts
- injury
- reporting an overly optimistic probability of achieving a pregnancy
- 125,000 deaths
- ovulation disorder
- conventional protocols of ovulation stimulation do not accommodate selective treatment regimens
- bone fracture(s)

Representative	text-small	text-large	Qwen3-4B	Qwen3-0.6B
10% admissions ~ non-compliance	5	8	2	4
age-related disabilities ~	5	4	4	3
infertility	2	11	7	19
difficulty in transporting the lifts	1	3	1	1
injury	2	2	4	6
reporting an overly	4	5	5	5
125,000 deaths	4	4	3	4
ovulation disorder	2	1	4	1
conventional protocols ~	1	2	4	1
bone fracture(s)	3	4	6	6

▲ 각 모델의 군집 개수

- 각 모델의 Representative 하위 데이터를 살펴보면, 대체로 각 데이터가 의미에 맞게 군집된 것을 확인할 수 있음. 다만, text-large 모델에서는 “125,000 deaths per year caused by non-compliance” 데이터가 “10% of admissions to regular hospitals due to non-compliance” 군집에 포함된 반면, text-small 모델에서는 “125,000 deaths” 군집에 포함되었음. 의미적으로 두 군집 모두에 해당 데이터가 들어갈 수 있는 것으로 판단. (나머지 모델에서는 해당 데이터가 선정된 10개 군집 중 어느 곳에도 포함되지 않음.)
- Representative가 “infertility”, “injury”와 같이 단일 단어로 짧게 표현된 경우, Qwen3-Embedding-0.6B 모델에서 군집의 개수가 많은 경향을 보임.
- “125,000 deaths”처럼 수치가 포함된 짧은 Representative의 경우, 각 모델의 군집 개수는 비슷하게 나타남. 반면, “10% of admissions to regular hospitals due to non-compliance”처럼 다소 긴 표현에서는 text-large 모델이 가장 많은 군집 개수를 보였으며, 의미적으로도 올바르게 군집화된 것으로 판단.