

Supplementary Material D

- Numeric data reported in Figure 2 (p. 2)
- Numeric data reported in Figure 3. (p. 3)

| Domain | Option | GPT-3.5 n (%) | GPT-4 n (%) | gBARD n (%) | GPT-4 vs GPT-3.5 vs gBARD |
|------------------------------|--------------------|------------------|----------------|----------------|----------------------------|
| Questions with Answers | Overall | 60 (100%) | 60 (100%) | 32 (53.3%) | <i>p-value = 6.193e-16</i> |
| | <i>Mechanisms</i> | 15 (100%) | 15 (100%) | 14 (93.3%) | |
| | <i>Indications</i> | 15 (100%) | 15 (100%) | 5 (33.3%) | |
| | <i>Toxicities</i> | 15 (100%) | 15 (100%) | 0 (0%) | |
| | <i>Prognosis</i> | 15 (100%) | 15 (100%) | 13 (86.7%) | |

| Domain | Option | GPT-3.5 n (%) | GPT-4 n (%) | gBARD n (%) | GPT-4 vs GPT-3.5 vs gBARD |
|-----------------|--------------------|------------------|----------------|----------------|----------------------------|
| Reproducibility | Overall | 53 (88.3%) | 57 (95%) | 16 (50%) | <i>p-value = 8.376e-07</i> |
| | <i>Mechanisms</i> | 13 (86.7%) | 15 (100%) | 6 (40%) | |
| | <i>Indications</i> | 13 (86.7%) | 15 (100%) | 1 (6.7%) | |
| | <i>Toxicities</i> | 13 (86.7%) | 14 (93.3%) | 0 (0%) | |
| | <i>Prognosis</i> | 14 (93.3%) | 13 (86.7%) | 9 (60%) | |

| Domain | Option | GPT-3.5 n (%) | GPT-4 n (%) | gBARD n (%) | GPT-4 vs GPT-3.5 vs gBARD |
|---------------|----------------------------|-------------------------|-----------------------|-----------------------|---------------------------|
| Accuracy | <i>Answer rated as "1"</i> | 4 (7.5%) | 0 (%) | 3 (18.8%) | <i>p-value = 0.03342</i> |
| | <i>Answer rated as "2"</i> | 18 (34%) | 14 (24.6%) | 6 (37.5%) | |
| | <i>Answer rated as "3"</i> | 31 (58.5%) | 43 (75.4%) | 7 (43.8%) | |
| Readability | <i>Answer rated as "1"</i> | 0 (0%) | 0 (0%) | 0 (0%) | <i>p-value = 0.02126</i> |
| | <i>Answer rated as "2"</i> | 1 (1.9%) | 0 (0%) | 2 (12.5%) | |
| | <i>Answer rated as "3"</i> | 52 (98.1%) | 57 (100%) | 14 (87.5%) | |
| Relevance | <i>Answer rated as "1"</i> | 0 (0%) | 0 (0%) | 1 (6.2%) | <i>p-value = 0.04107</i> |
| | <i>Answer rated as "2"</i> | 12 (22.6%) | 16 (28.1%) | 8 (50%) | |
| | <i>Answer rated as "3"</i> | 41 (77.4%) | 41 (71.9%) | 7 (43.8%) | |