

日本産業動物獣医療における大規模言語モデルの 診断能力に関する評価指標の研究

岡山理科大学 理工学研究科 房ハルノ

- 目的** 日本の産業動物獣医療におけるLLMの診断性能を体系的に評価し、今後のモデル開発のベースラインを把握すること。
- 内容** 産業動物臨床学QA（知識量評価）と産業動物臨床Clinic（臨床推論能力評価）の2種類のベンチマークを提案し、それらに基づいて既存モデルの性能を評価するとともに、得られた結果に基づいて詳細な分析を行った。
- 結果** 一部のオープンソースモデル、例えばLlama-4-Scoutは、臨床知識評価では79.2%の正答率を示した一方で、対話型臨床推論評価では6.7%にとどまったことから、診断性能の低さは単なる専門知識の不足だけでなく、より本質的には対話的臨床推論能力の制約に起因する可能性が示唆された。

利用した計算機 GPUノード群

ノード時間

513時間