Length of Procedure and F1 Score

In the case of GPT-3, model displayed an obvious performance discrepancy between short and long procedures for tracking entity state and answer event likelihood questions.

In Codex, such discrepancy also exist but is much more nuanced than that of GPT-3

Codex Performance on answering event likelihood questions for procedures with below average number of steps:

F1 Score = 0.63

Accuracy = 0.82

Codex performance on answering event likelihood questions for procedures with above average number of steps:

F1 Score= 0.67

Accuracy = 0.85