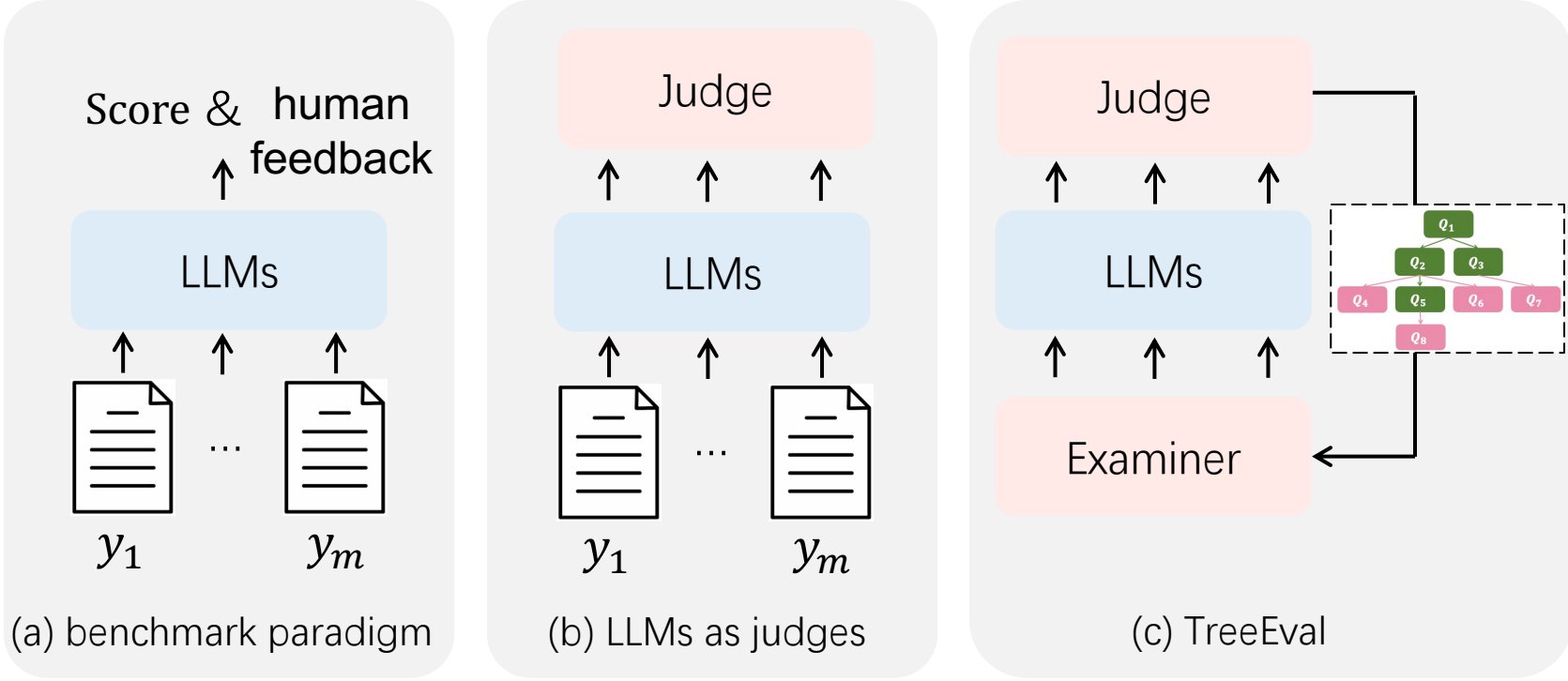
Score & human feedback
LLMs
y_1
y_m
(a) benchmark paradigm
Judge
LLMs
y_1
y_m
(b) LLMs as judges
Judge
LLMs
Examiner
Q_1
Q_2
Q_3
Q_4
Q_5
Q_6
Q_7
Q_8
(c) TreeEval