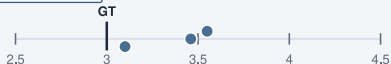


(a) Absolute calibration

$$G(p, g) = \exp\left(-\frac{(p-g)^2}{2\sigma^2}\right), \sigma = 0.5$$



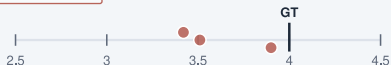
Group A · g=3.00



sample	prediction	Rcal
A1	3.55	0.55
A2	3.46	0.65
A3	3.10	0.98



Group B · g=4.00



sample	prediction	Rcal
B1	3.42	0.51
B2	3.51	0.62
B3	3.90	0.90

Closer predictions receive higher pointwise reward.

(b) Cross-group ranking



A · g=3.00

A < B



B · g=4.00



pair	predicted order	status	Rrank
A1-B1	3.55 > 3.42	reversed	-0.13
A2-B2	3.46 < 3.51	correct	+0.05
A3-B3	3.10 < 3.90	correct	+0.66

Only valid cross-group pairs with the same specialist dimension and a sufficient reference-score gap δ are compared.

Combined reward

calibration

+

ranking

+

semantic

+

format