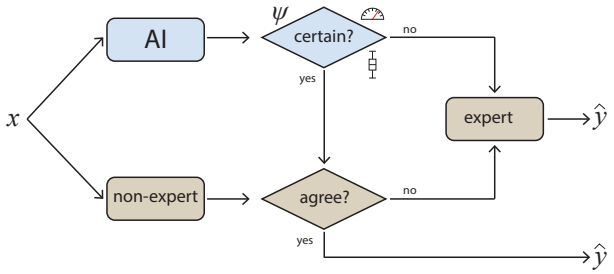

Fig. 5. Diagram of arbitration reading. For the standalone human system, a non-expert replaces the AI and certainty estimator.

Table 2. Model performance on InD, OoD-H and OoD-D datasets measured by F1, Brier, NLL, ECE, AURC and AUC_{mis}. While these traditional metrics show variation in model performances, our summary pAUJF1C metric shows similar level of clinical performance when used to complement doctors.

Method	F1 ↑	Brier ↓	NLL ↓	ECE ↓	AURC ↓	AUC_{mis} ↑	$pAUJF1C_{0.5}$ ↑	$pAUJF1C_{0.75}$ ↑	$pAUJF1C_{0.9}$ ↑
In-distribution (InD)									
MLE	**0.793**	0.126	0.410	0.047	0.079	0.727	**0.394**	0.195	**0.075**
TTA	0.789	0.130	0.422	0.065	0.076	0.732	0.393	0.195	0.074
TS	0.787	0.128	0.409	0.053	0.077	0.740	**0.394**	**0.196**	**0.075**
MCD	0.759	0.128	0.411	0.043	0.080	0.771	0.393	0.193	0.073
SVGD	0.741	0.130	0.417	0.051	0.085	**0.798**	0.388	0.190	0.071
SWAG	0.783	**0.121**	**0.388**	0.060	**0.067**	0.773	0.390	0.194	0.074
Ensemble	0.787	0.123	0.392	**0.027**	0.068	0.747	0.390	0.191	0.073
Out-of-distribution Hospital (OoD-H)									
MLE	0.723	0.150	0.452	0.039	0.092	0.760	0.368	0.180	0.069
TTA	0.717	0.161	0.489	0.073	0.10	0.746	0.364	0.178	0.068
TS	0.723	0.148	0.451	0.041	0.092	0.760	0.368	0.180	0.068
MCD	0.732	0.155	0.470	0.070	0.102	0.719	0.366	0.180	0.069
SVGD	0.729	0.143	0.442	**0.032**	0.087	**0.781**	0.374	0.183	0.069
SWAG	0.717	0.150	0.466	0.057	0.090	**0.781**	0.369	0.179	0.068
Ensemble	**0.760**	**0.142**	**0.428**	0.037	**0.079**	0.752	**0.380**	**0.186**	**0.071**
Out-of-distribution Device (OoD-D)									
MLE	0.762	0.136	0.421	0.033	0.080	0.754	0.390	0.193	0.073
TTA	0.759	0.141	0.449	0.039	0.089	0.741	**0.392**	**0.194**	0.073
TS	0.762	0.139	0.435	0.056	0.080	0.754	0.390	0.192	0.073
MCD	0.747	0.137	0.424	0.025	0.081	0.763	0.388	0.191	0.072
SVGD	0.750	0.135	0.423	**0.018**	0.081	0.764	0.389	0.193	0.073
SWAG	**0.764**	**0.132**	**0.409**	0.033	**0.072**	**0.770**	0.391	0.192	0.074
ENSEMBLE	0.760	0.135	0.419	0.023	0.077	0.764	0.389	0.192	**0.076**

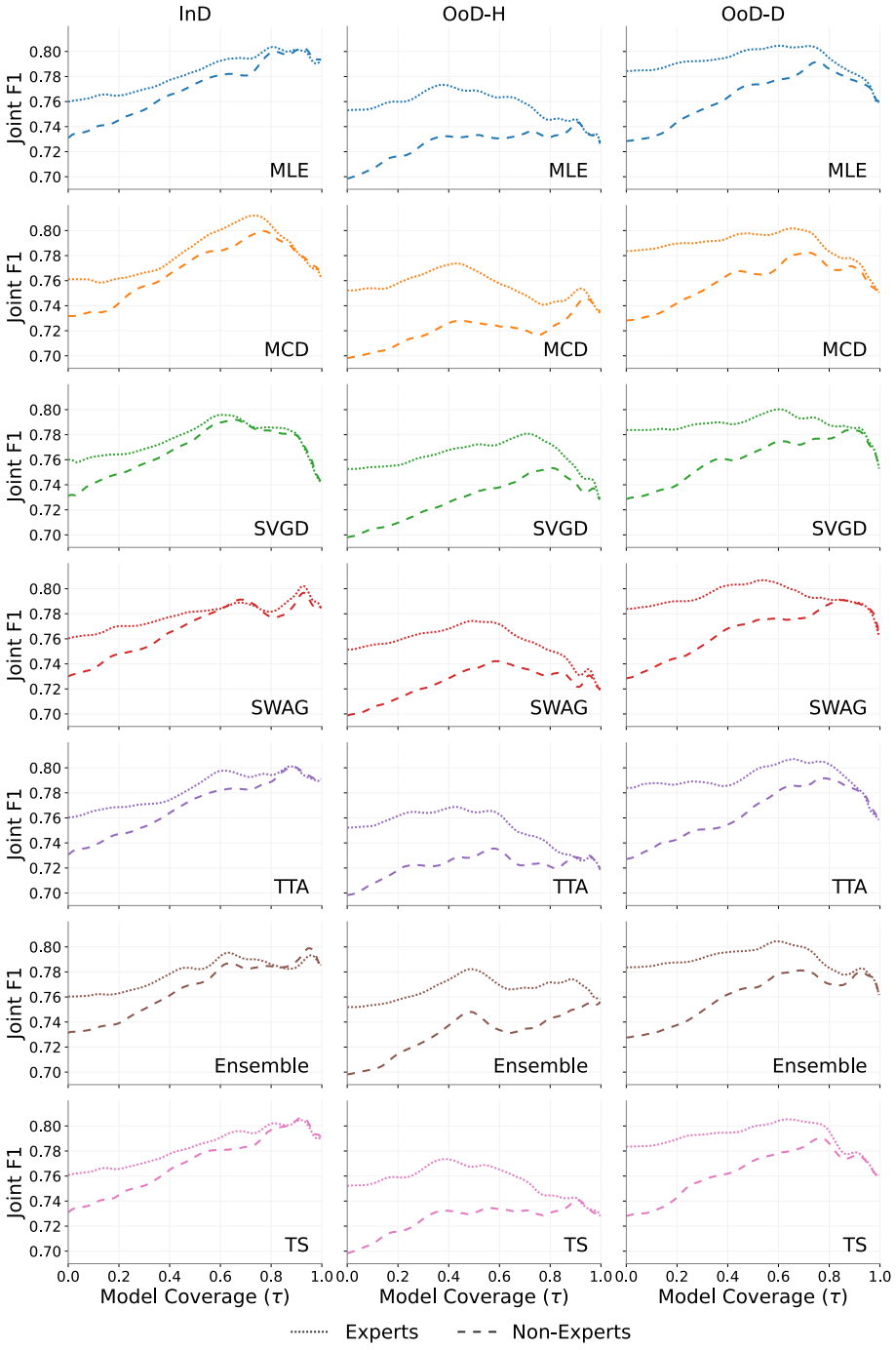

Fig. 6. Joint F1 coverage curves reveal that human performance significantly impacts the joint system performance, particularly at lower model coverage.