Appendix

Limitations and Future Work

Feature Sensitivity. The current feature vector ϕ_{ij} relies heavily on pre-trained BERT embeddings. While effective, this design assumes that semantic similarity is a *sufficient* proxy for self-preference. Future work could incorporate *instruction-level* features (e.g., task category, prompt difficulty) or *meta-judge* signals (e.g., perplexity under the judge's own decoding) to better disentangle intrinsic quality from stylistic preference.

Objective Tradeoffs. The consensus loss $\mathcal{L}$ treats all judges as *equally informative*, which may *over-correct* high-quality judges like GPT-4o (Figure 4). A weighted variant that down-weights judges with historically poor correlation could preserve *discriminative power* while still reducing bias.

Scalability. Training the adapter on 450K pairwise judgments is computationally lightweight (8 hours on a single 4090), but the $O(MN)$ complexity of the consensus loss becomes prohibitive for $M \gg 100$. Approximate solutions—e.g., clustering judges into *meta-judges*—could retain debiasing benefits at scale.

Additional Empirical Analysis

Loss-driven Convergence. Recall that the training objective (Eq. 3) enforces three complementary desiderata: (i) *anchor alignment* via MSE, (ii) *rank fidelity* via Pearson correlation, and (iii) *collective consensus* via global MSE. Table 4 shows the zero-shot transfer performance on the HUMAN-ANNOTATED TRANSFER SET. UDA boosts the mean Pearson correlation with the baseline-Elo consensus from 0.7363 to 0.8778 and slashes the average squared distance to the anchor scores from $43,152.6$ to $19,452.7$. This confirms that the loss composition successfully suppresses judge-specific noise while maintaining fidelity to a stable, judge-agnostic ranking.

Prompts for Answer Generation and Judging

We release *all* prompts used to (i) elicit model answers and (ii) prompt the LLM judges. The answer prompt and judge prompt are fixed across models to ensure reproducibility. These prompts originate from our empirical observations of and iterative refinements to both the large language model's responses and the pairwise evaluation procedure.

Answer Generation Prompt

System: You are the most advanced artificial intelligence ever created by humankind, possessing extensive knowledge of the world. Please answer the following question:
User: {prompt}

Model	Pearson ↑		MSE ↓	
	baseline	UDA	baseline	UDA
gpt-4o	0.9432	0.9280	14070.3	13697.7
deepseek-v3	0.9102	0.9868	11559.5	13369.0
claude-3.5	−0.2248	0.1791	108549.5	66849.0
glm4-plus	0.9746	0.9614	59878.7	6080.4
glm4-air	0.8971	0.9627	32496.8	9160.3
glm4-flash	0.1865	0.8890	72145.9	34508.7
doubao-1.5pro	0.9066	0.9906	25488.3	8438.5
qwen-max	0.8955	0.9629	19261.0	22277.0
gemini-2.0-flash	0.9222	0.9586	52760.4	9481.8
deepseek-r1	0.9518	0.9590	35315.3	10664.2
Average	0.7363	**0.8778**	43152.6	**19452.7**

Table 4: Analysis of Alignment with Baseline Consensus. Zero-shot transfer results on HUMAN-ANNOTATED TRANSFER SET. UDA improves average Pearson correlation by $+0.1415$ and reduces average MSE by $23{,}699.9$ without retraining. We adopt the baseline's mean scores as ground-truth supervision.

Pairwise Judging Prompt

System: Please act as an impartial adjudicator and assess the quality of the two AI assistants' responses to the user's query below. Your evaluation should adopt the user's perspective and consider dimensions such as the assistants' helpfulness.
User: Scoring Criteria 1. Compare the quality of the two responses in relation to the user's query and provide a concise justification. Avoid positional bias; the order of presentation must not influence your decision. 2. If the question is a highly specialized, objective factual query that you cannot accurately adjudicate, return [[C]] to prevent misinformation. 3. If either response contains errors, explicitly identify them. Do not endorse mutually contradictory statements as simultaneously correct (i.e., do not mark opposing claims from A and B as both being right). 4. Output exactly one of: [[A]] – Assistant A's response is of higher quality. [[B]] – Assistant B's response is of higher quality. [[C]] – The two responses are of equal quality.
Additional Notes - Begin by checking for factual accuracy; if an error is present, prioritize it in your judgment. - Scores must be based on objective criteria, not subjective preference. - Response length is irrelevant; evaluate whether the content genuinely addresses the user's need from the user's perspective. - Evaluate quality strictly from the user's standpoint. **Example** User question: What is the English term for ”Mining transportation and maintenance fees”? Assistant A: The English term is “maintenance fee,” which generally denotes the cost required to keep facilities or equipment in good working condition, including repairs, upkeep, and related expenses. Assistant B: The correct English term is “Mine transportation maintenance fee.” Analysis: Assistant A's response contains a factual error; the accurate translation is “Mine transportation maintenance fee.” Assistant B's answer is correct and satisfies the user's requirement. Result: [[B]]
Return Format
Analysis: . . . Result: [[x]]
Input
[User Question – Start]
question
[User Question – End]
[Assistant A – Start]
answer_i
[Assistant A – End]
[Assistant B – Start]
answer_j

Proofs

Complete Proof of Aggregate Bias Reduction

Mathematical Setup. We analyze an idealized mathematical model of the consensus alignment process to formally demonstrate how it reduces aggregate bias. Let $\mathcal{J} = \{1, 2, \ldots, M\}$ be the set of judges. For a given model m, let R_m^* be its unknown true Elo score. When evaluated by judge $k \in \mathcal{J}$, the resulting score is

$$R_m^{(k)}.$$

We define the bias of judge k for model m as

$$\epsilon_m^{(k)} = R_m^{(k)} - R_m^*.$$

This bias $\epsilon_m^{(k)}$ can be positive (over-rating, e.g., *glm-4-flash*) or negative (under-rating, e.g., *gpt-4o*), reflecting the self-preference patterns observed in our experiments.

The baseline Elo system produces scores $R_{m,\text{base}}^{(k)}$ with biases $\epsilon_{m,\text{base}}^{(k)}$. The UDA framework updates these scores by pulling them towards a consensus. Let the post-alignment score for judge k be $R_{m,\text{UDA}}^{(k)}$. The objective function in Eq. (3) effectively models the updated score as a shrinkage towards the mean score:

$$R_{m,\text{UDA}}^{(k)} \approx \alpha R_{m,\text{base}}^{(k)} + (1-\alpha)\bar{R}_{m,\text{base}},$$

where

$$\bar{R}_{m,\text{base}} = \frac{1}{M}\sum_{j=1}^{M} R_{m,\text{base}}^{(j)}$$

is the average (consensus) score across all judges, and $\alpha \in [0, 1]$ is a parameter learned by the network that determines the strength of this alignment. The closer a judge is to the consensus, the larger the effective α.

Post-Alignment Bias. The new bias for judge k after UDA is

$$\epsilon_{m,\text{UDA}}^{(k)} = R_{m,\text{UDA}}^{(k)} - R_m^*.$$

Substituting $R_{m,\text{base}}^{(k)} = R_m^* + \epsilon_{m,\text{base}}^{(k)}$ and $\bar{R}_{m,\text{base}} = R_m^* + \bar{\epsilon}_{m,\text{base}}$ (with $\bar{\epsilon}_{m,\text{base}} = \frac{1}{M}\sum_{j=1}^{M}\epsilon_{m,\text{base}}^{(j)}$), we obtain

$$\begin{aligned}
\epsilon_{m,\text{UDA}}^{(k)} &= \alpha\big(R_m^* + \epsilon_{m,\text{base}}^{(k)}\big) + (1-\alpha)\big(R_m^* + \bar{\epsilon}_{m,\text{base}}\big) - R_m^* \\
&= \alpha\epsilon_{m,\text{base}}^{(k)} + (1-\alpha)\bar{\epsilon}_{m,\text{base}}.
\end{aligned}$$

Aggregate Absolute Bias Reduction. Let

$$S_\text{base} = \sum_{k=1}^{M}\big|\epsilon_{m,\text{base}}^{(k)}\big|, \qquad S_\text{UDA} = \sum_{k=1}^{M}\big|\epsilon_{m,\text{UDA}}^{(k)}\big|$$

be the total absolute biases before and after UDA.

Theorem .2. *For any set of biases $\{\epsilon_{m,base}^{(k)}\}_{k=1}^{M}$ and for any $\alpha \in [0, 1]$, the total absolute bias after consensus alignment does not increase:*

$$S_{UDA} \le S_{base}.$$

Proof. By the triangle inequality on each term of S_UDA,

$$\begin{aligned}
\big|\epsilon_{m,\text{UDA}}^{(k)}\big| &= \big|\alpha\epsilon_{m,\text{base}}^{(k)} + (1-\alpha)\bar{\epsilon}_{m,\text{base}}\big| \\
&\le \alpha\big|\epsilon_{m,\text{base}}^{(k)}\big| + (1-\alpha)\big|\bar{\epsilon}_{m,\text{base}}\big|.
\end{aligned}$$

Summing over all judges,

$$\begin{aligned}
S_\text{UDA} &\le \sum_{k=1}^{M}\Big(\alpha\big|\epsilon_{m,\text{base}}^{(k)}\big| + (1-\alpha)\big|\bar{\epsilon}_{m,\text{base}}\big|\Big) \\
&= \alpha S_\text{base} + (1-\alpha)M\big|\bar{\epsilon}_{m,\text{base}}\big|.
\end{aligned}$$

Applying Jensen's inequality to the convex function $|\cdot|$,

$$\big|\bar{\epsilon}_{m,\text{base}}\big| = \Big|\frac{1}{M}\sum_{j=1}^{M}\epsilon_{m,\text{base}}^{(j)}\Big| \le \frac{1}{M}\sum_{j=1}^{M}\big|\epsilon_{m,\text{base}}^{(j)}\big| = \frac{S_\text{base}}{M}.$$

Substituting this bound,

$$S_\text{UDA} \le \alpha S_\text{base} + (1-\alpha)M\frac{S_\text{base}}{M} = S_\text{base}.$$

Hence, the total absolute bias is not increased. ■

Conclusion. The UDA procedure is guaranteed to not increase (and in practice, strictly reduce, unless all biases are already identical) the total absolute bias across the system of judges. By pulling extreme outliers—both positive and negative—towards a central consensus, it produces a more stable and less biased overall ranking, which is validated by the significant reduction in inter-judge score standard deviation and improved correlation with human judgments observed in our experiments.

Full judge score Matrices

Notation. Each row corresponds to one judge model; each column corresponds to one answer model. Models are ordered as: `gpt-4o`, `deepseek-v3`, `claude-3.5`, `glm-4-plus`, `glm-4-air`, `glm-4-flash`, `doubao-1.5pro`, `qwen-max`, `gemini-2.0-flash`, `deepseek-r1`.

Full Judge Score Matrices for Ablation Study We provide the complete judge-score matrices from the ablation study conducted on the Human-Annotated Transfer Set. Each row corresponds to a judge, and each column corresponds to a candidate model. The models are ordered as listed in the main paper.

Reproducibility Checklist. Code, model checkpoints and the Human-Annotated Transfer Set are released at

https://anonymous.4open.science/r/62AB93CD-23B4

Judge	gpt-4o	deepseek-v3	claude-3.5	glm-4-plus	glm-4-air	glm-4-flash	doubao-1.5pro	qwen-max	gemini-2.0-flash	deepseek-r1
J1	1123.98	1314.97	1339.64	1092.25	960.67	726.34	1180.35	1389.14	1280.25	1592.40
J2	1344.56	1507.43	1245.94	957.63	774.87	693.88	1254.53	1301.13	1163.73	1756.31
J3	1430.04	1519.86	1504.93	1030.06	866.62	597.91	1373.22	983.95	820.78	1872.64
J4	1016.41	1201.16	1196.69	1184.40	945.62	712.98	1027.81	1374.29	1702.74	1637.88
J5	1073.32	1234.74	1130.96	1190.02	967.96	807.96	1062.67	1405.89	1633.85	1492.63
J6	1146.40	1451.35	1133.97	1022.55	960.62	949.84	1172.75	1364.08	1321.84	1476.60
J7	1156.73	1282.67	1317.72	1102.28	704.31	463.17	1250.19	1420.78	1731.67	1570.48
J8	1176.31	1574.78	1384.54	913.84	790.57	546.31	1076.23	1453.41	1237.85	1846.17
J9	873.45	1110.10	1411.40	1053.64	826.41	526.93	1220.16	1374.48	1991.27	1612.16
J10	1006.40	1401.21	1346.90	911.50	632.07	383.50	1283.69	1561.84	1568.52	1904.37
UDA	1011.84	1250.88	1407.04	1103.99	911.00	825.96	1208.18	1356.61	1379.88	1544.61
UDA	1113.52	1323.18	1341.04	1012.35	872.24	809.57	1203.75	1363.79	1279.60	1680.96
UDA	1216.48	1394.90	1370.49	1066.91	833.52	792.63	1277.65	1262.71	1181.19	1603.51
UDA	1032.41	1326.42	1406.93	941.68	839.66	819.68	1088.63	1430.47	1549.08	1565.04
UDA	1112.89	1311.94	1370.03	1038.88	759.82	876.91	1088.26	1392.70	1553.80	1494.76
UDA	1158.16	1386.46	1309.95	1062.83	941.13	846.27	1149.40	1333.55	1344.87	1467.37
UDA	1051.69	1373.38	1329.12	1038.44	817.83	676.24	1148.93	1455.43	1517.84	1591.10
UDA	1124.66	1376.73	1354.04	1040.37	844.24	713.35	1178.46	1400.94	1290.47	1676.75
UDA	979.56	1321.11	1476.07	1027.03	853.74	791.86	1207.69	1347.74	1460.67	1534.51
UDA	1096.99	1370.99	1392.02	1028.23	824.30	663.20	1146.43	1463.64	1490.27	1523.94

Table 5: **ArenaHard Dataset: baseline Elo (top) vs. UDA Method (bottom)**

Judge	gpt-4o	deepseek-v3	claude-3.5	glm-4-plus	glm-4-air	glm-4-flash	doubao-1.5pro	qwen-max	gemini-2.0-flash	deepseek-r1
J1	820.93	1169.77	1063.86	1180.98	912.45	746.43	1289.74	1484.22	1454.52	1877.10
J2	1281.40	1309.34	1078.76	1210.14	959.14	664.53	1190.17	1402.90	1326.95	1576.69
J3	1459.42	1434.81	1148.82	1341.07	1154.05	1082.81	1158.53	1110.92	1158.02	951.53
J4	727.93	1375.24	895.98	1122.83	710.39	523.17	1350.34	1641.94	1405.05	2247.13
J5	698.44	1184.78	1060.97	1080.97	876.92	866.45	1233.04	1633.95	1344.93	2019.54
J6	1387.21	1374.16	1198.25	1235.12	1020.42	1117.67	1061.98	1388.73	1103.78	1112.69
J7	852.40	1452.30	1114.15	1309.11	755.75	613.18	1636.28	1601.80	1106.17	1558.86
J8	1143.36	1324.79	1097.86	1280.63	893.90	976.86	1311.58	1276.34	1332.57	1362.10
J9	606.89	1191.82	1212.54	1035.90	797.55	595.32	1532.11	1478.94	1432.74	2116.16
J10	1197.16	1508.89	1020.43	1220.23	591.70	361.67	1506.17	1442.62	1346.96	1804.17
UDA	929.83	1221.50	1105.68	1212.09	1044.76	976.05	1297.16	1364.22	1348.20	1500.51
UDA	1135.44	1266.64	1095.96	1199.45	998.51	948.20	1264.47	1370.28	1283.76	1437.28
UDA	1312.28	1272.16	992.62	1238.63	1208.68	1126.88	1290.37	1170.10	1199.25	1189.03
UDA	958.25	1247.67	1080.86	1118.82	953.30	923.91	1318.88	1452.59	1350.77	1594.96
UDA	976.49	1227.70	1160.61	1223.52	942.34	969.75	1264.26	1381.11	1306.87	1547.33
UDA	1191.71	1246.18	1194.63	1231.78	1102.26	1004.05	1162.83	1293.45	1265.00	1307.13
UDA	1030.98	1272.61	1156.56	1208.91	975.73	925.70	1311.05	1334.97	1266.51	1516.99
UDA	1179.94	1300.77	1111.39	1197.34	1037.00	989.38	1234.67	1280.93	1279.26	1389.31
UDA	986.95	1239.54	1121.67	1179.81	1017.57	946.45	1354.42	1394.33	1197.81	1561.45
UDA	1124.61	1303.36	1069.66	1151.73	971.80	889.36	1381.52	1362.55	1311.55	1433.87
Human	1049.80	1284.72	663.97	1386.02	927.27	717.69	1329.97	1485.66	1315.73	1839.18

Table 6: **Human-Annotated Transfer Set: baseline Elo (top), UDA Method (mid), and Human score (bottom)**

Judge	gpt-4o	d-seek-v3	claude-3.5	glm4-plus	glm4-air	glm4-flash	doubao-pro	qwen-max	gemini-flash	d-seek-r1
Baseline Elo Scores										
J1	820.93	1169.77	1063.86	1180.98	912.45	746.43	1289.74	1484.22	1454.52	1877.10
J2	1281.40	1309.34	1078.76	1210.14	959.14	664.53	1190.17	1402.90	1326.95	1576.69
J3	1459.42	1434.81	1148.82	1341.07	1154.05	1082.81	1158.53	1110.92	1158.02	951.53
J4	727.93	1375.24	895.98	1122.83	710.39	523.17	1350.34	1641.94	1405.05	2247.13
J5	698.44	1184.78	1060.97	1080.97	876.92	866.45	1233.04	1633.95	1344.93	2019.54
J6	1387.21	1374.16	1198.25	1235.12	1020.42	1117.67	1061.98	1388.73	1103.78	1112.69
J7	852.40	1452.30	1114.15	1309.11	755.75	613.18	1636.28	1601.80	1106.17	1558.86
J8	1143.36	1324.79	1097.86	1280.63	893.90	976.86	1311.58	1276.34	1332.57	1362.10
J9	606.89	1191.82	1212.54	1035.90	797.55	595.32	1532.11	1478.94	1432.74	2116.16
J10	1197.16	1508.89	1020.43	1220.23	591.70	361.67	1506.17	1442.62	1346.96	1804.17
UDA (Full) Scores										
J1	1141.36	1202.67	1138.28	1185.60	1107.94	1116.15	1196.94	1255.72	1311.88	1343.46
J2	1336.81	1233.62	1150.74	1161.55	1110.13	1057.50	1180.80	1192.78	1299.93	1276.15
J3	1377.93	1280.68	1105.90	1268.59	1148.17	1192.59	1221.75	1141.70	1215.67	1047.01
J4	1196.71	1175.79	1084.40	1183.14	1079.98	1038.87	1235.12	1290.29	1267.97	1447.74
J5	1076.92	1195.13	1152.85	1157.65	1079.05	1142.77	1217.09	1291.34	1271.03	1416.17
J6	1305.32	1283.27	1193.84	1197.23	1148.34	1184.03	1158.19	1209.66	1211.69	1108.43
J7	1144.90	1250.18	1200.55	1210.39	1083.73	1097.62	1371.63	1232.65	1094.98	1313.39
J8	1197.07	1237.31	1183.12	1190.55	1068.54	1157.58	1232.84	1194.02	1335.07	1203.90
J9	1100.04	1165.76	1212.91	1158.41	1160.29	1064.05	1252.41	1262.05	1136.86	1487.23
J10	1241.42	1252.49	1124.28	1156.55	1034.83	1007.34	1323.65	1243.34	1258.30	1357.81
UDA (Ablated) Scores – w/o Self-Features										
J1	1102.51	1175.33	1150.11	1158.92	1021.45	998.64	1201.20	1308.81	1244.17	1441.86
J2	1188.19	1210.87	1162.30	1182.25	1055.78	1010.99	1195.81	1300.15	1230.90	1403.76
J3	1235.40	1244.11	1120.57	1211.70	1088.31	1090.25	1188.93	1255.44	1222.01	1313.28
J4	1080.15	1221.90	1098.88	1140.33	999.50	965.87	1210.74	1340.22	1250.60	1511.81
J5	1055.77	1180.25	1145.92	1133.19	1011.66	1022.40	1205.11	1335.98	1240.15	1469.57
J6	1210.93	1233.50	1166.14	1190.88	1066.02	1088.71	1180.55	1280.60	1220.31	1322.36
J7	1100.28	1255.12	1177.36	1195.40	1015.99	988.54	1245.33	1322.71	1199.88	1449.39
J8	1155.61	1201.78	1160.01	1177.66	1033.47	1060.11	1199.95	1277.30	1255.13	1378.98
J9	1044.89	1177.01	1188.45	1135.22	1050.18	990.56	1225.10	1318.88	1215.99	1453.72
J10	1150.36	1241.92	1118.04	1165.01	1010.96	948.79	1228.69	1315.52	1255.71	1414.00

Table 7: Full judge-score matrices for Baseline Elo, the full UDA method, and the ablated UDA variant on the HUMAN-ANNOTATED TRANSFER SET. The ablated model shows visibly lower column-wise variance but diverges more from human-aligned rankings.