A.4 DATA CONSTRUCTION PIPELINE

Table A1: Data Construction Pipeline: Step-by-Step Example

Step	Component	Example Content
Step 1	Target Entity A	*Hornussen (Swiss team striking sport)*
	Distractor B	*Baseball (globally popular team bat-and-ball sport)*
Step 2	Shared Attributes	Team-based striking game; offense/defense alternation; players take turns hitting; projectiles reach very high speeds ($>$100 mph); protective gear required; governed by formal associations or leagues.
	Distinctive Attributes	**Hornussen:** strikes a plastic puck ("Nouss") with whip-like swing using a long wooden rod; defenders intercept with wooden boards; fan-shaped field $\sim$300m; 18–20 defenders spread in wide formation; scoring depends on distance/landing point.
Step 3	Ambiguous Question Q	"Which team-based striking sport features two sides alternating offense and defense, where individuals sequentially hit a high-speed projectile and teammates coordinate to intercept it in the air? Outcomes depend on whether the projectile is intercepted or lands within the valid playing field. Defense relies on wide positioning and collaboration, all offensive players take turns striking, flight speeds often exceed 100 mph, protective gear is required due to impact risk, and the sport is governed by long-standing associations or leagues."
Step 4	Contextual Information	– Struck object is a plastic puck, resembling an ice hockey puck. – Striking method uses a whip-like swing with a long wooden rod. – Defenders use wooden boards to block the puck in mid-air. – Field: fan shape, $\sim$300m long, 10–12° angle. – Defensive line: 18–20 players. – Scoring: distance/landing-based.
Step 5	Reasoning Path	Q gives a plausible candidate set (e.g., Baseball vs Hornussen). Adding context clarifies unique Hornussen features (puck, whip swing, fan-shaped field, defensive boards), leading to the unique answer = Hornussen.

A.5 ROUNDS CONSTRAINTS RESULTS

Table A2: Performance comparison of models under varying average interaction levels. Metrics include accuracy (%), expected calibration error (ECE, %), and average rounds of interaction (Interaction).

Interaction	Accuracy (%)	ECE (%)	Setting
GPT-5			
1.14	14.0	71.50	SCALING
1.76	16.0	71.54	SCALING
1.90	20.0	70.06	SCALING
4.40	24.0	63.34	FORCED
6.10	20.0	69.02	FORCED
8.32	32.0	54.68	FORCED
9.40	40.0	48.20	FORCED
11.56	40.0	46.86	FORCED
DeepSeek-Chat			
0.38	10.0	77.00	SCALING
0.74	8.0	80.30	SCALING
1.54	10.0	75.20	SCALING
5.40	20.0	62.30	FORCED
6.68	22.0	52.60	FORCED
9.26	22.0	61.30	FORCED
10.82	16.0	66.40	FORCED
12.62	24.0	61.20	FORCED
Claude-Sonnet-4			
0.16	6.0	79.90	SCALING
0.70	4.0	80.24	SCALING
0.78	8.0	81.84	SCALING
3.12	12.0	75.80	FORCED
4.26	12.0	76.90	FORCED
6.10	10.0	76.00	FORCED
8.02	16.0	69.10	FORCED
10.46	18.0	68.40	FORCED
GPT-4o-mini			
2.00	4.0	49.50	SCALING
3.62	8.0	47.60	SCALING
2.76	8.0	33.20	SCALING
14.50	4.0	65.50	FORCED
16.50	4.0	69.70	FORCED
14.88	2.0	62.60	FORCED
11.18	6.0	56.10	FORCED
14.92	2.0	66.70	FORCED