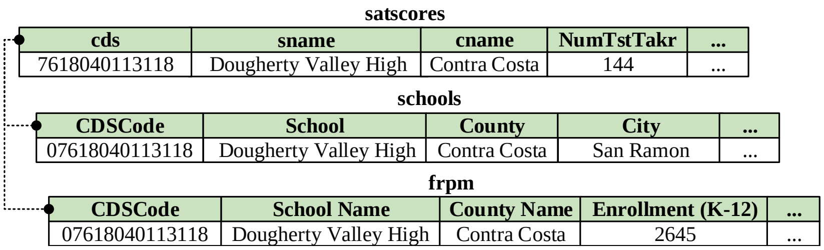
Case Study

Table I illustrates a case study of the predictions generated by baselines with CLEAR on BIRD. For query ambiguity, the parser considers “each” to “each superhero individually”, thus stating `SELECT superhero_name` and `GROUP BY superhero_name`. However, the correct prediction is generated through the discovery and intent clarification by CLEAR. Regarding match ambiguity, the parser chooses `schools.School` and `schools.County` when aligning the NL to the schema due to the relevant semantics of words. However, the actual requirement of the gold is ambiguous columns, which is identified and corrected by CLEAR.

TABLE I: Case study of CLEAR on BIRD.

	Query Ambiguity
Question:	*Calculate the average height **for each superhero**.*
Gold SQL:	SELECT CAST(SUM(height_cm) AS REAL) / COUNT(id) FROM superhero
MAC-SQL: **(Incorrect)**	SELECT superhero_name, AVG(height_cm) AS average_height FROM superhero WHERE height_cm IS NOT NULL GROUP BY superhero_name
MAC-SQL: **+ CLEAR** **(correct)**	φ: {for each superhero: [for each superhero individually, common to all superheroes]} ψ: {for each superhero: common to all superheroes} $\tilde{Q}$: Calculate the average height common to all superheros. **SQL**: SELECT SUM(height_cm) / COUNT(*) AS average_height FROM superhero WHERE height_cm IS NOT NULL
	Match Ambiguity
Question: Schema:	*Which **school** in **Contra Costa** has the highest number of test takers?* satscores: sname, cname, NumTstTakr | schools: School, County
Gold SQL:	SELECT sname FROM satscores WHERE cname = ‘Contra Costa’ AND sname IS NOT NULL ORDER BY NumTstTakr DESC LIMIT 1
MAC-SQL: **(Incorrect)**	SELECT T2.School FROM satscores AS T1 INNER JOIN schools AS T2 ON T1.cds = T2.CDSCode WHERE T2.County = ‘Contra Costa’ AND T1.NumTstTakr IS NOT NULL GROUP BY T2.School ORDER BY MAX(T1.NumTstTakr) DESC LIMIT 1
MAC-SQL: **+ CLEAR** **(correct)**	φ: {school: [{schools: [School]}, {satscores: [sname]}], Contra Costa: [{schools: [County]}, {satscores: [cname]}]} ψ: {school: {satscores: [sname]}, Contra Costa: {satscores: [cname]}} $\tilde{Q}$: Which school (satscores.sname) in Contra Costa (satscores.cname) has the highest number of test takers? **SQL**: SELECT sname FROM satscores WHERE cname = ‘Contra Costa’ AND sname IS NOT NULL ORDER BY NumTstTakr DESC LIMIT 1



satscores

cds	sname	cname	NumTstTakr	...
7618040113118	Dougherty Valley High	Contra Costa	144	...

schools

CDSCode	School	County	City	...
07618040113118	Dougherty Valley High	Contra Costa	San Ramon	...

frpm

CDSCode	School Name	County Name	Enrollment (K-12)	...
07618040113118	Dougherty Valley High	Contra Costa	2645	...

Fig. 1: Example of schema on BIRD.