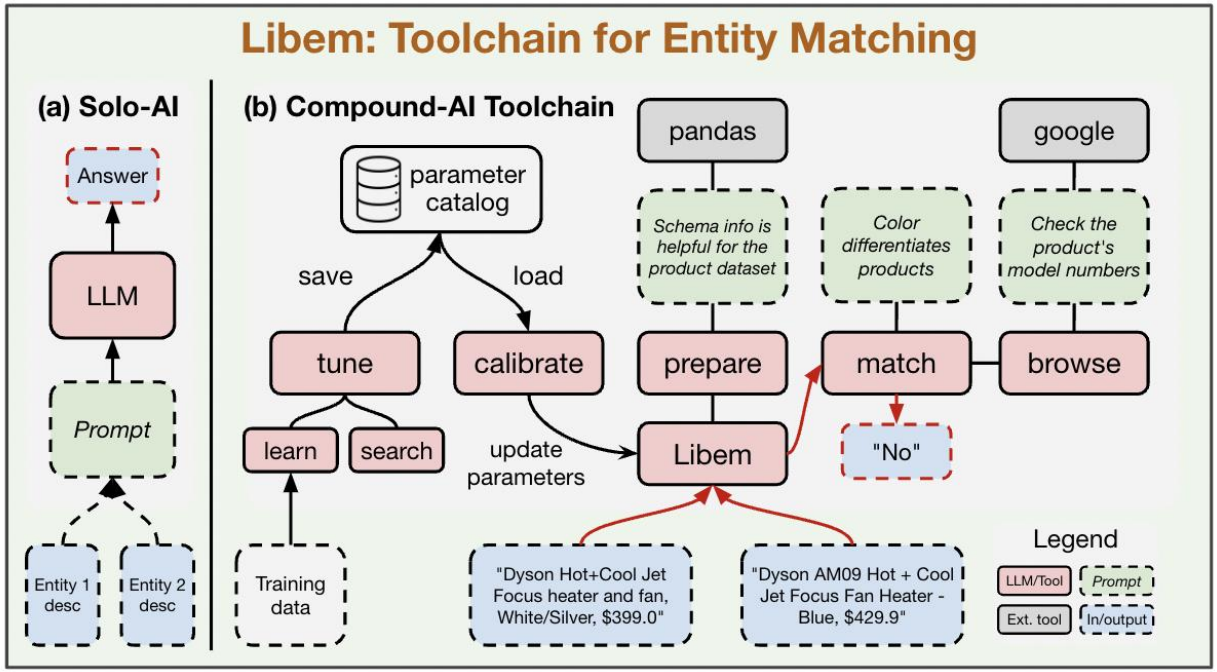
# Liberal Entity Matching as a Compound AI Toolchain

Silvery Fu, David Wang, Kathleen Ge, Wen Zhang
UC Berkeley, System Design Studio

**Task: Entity Matching.** Determine whether two descriptions refer to the same entity.




**Today**: Solo-AI EM relies on hand-tuned prompts and static knowledge.

**Approach**: Compound-AI EM that enhances LLMs with tools and optimizations.

arena.libem.org

# Compound Schema Registry

Silvery Fu, Xuewei (Sylvia) Chen
UC Berkeley, System Design Studio

## Schema Transformation Language (STL)

| Command class | Command name | Description |
|---|---|---|
| Schema matching | MATCH | Used to determine whether the source and target schemas **correspond to the same entity; if they match**, the schema mapping will **continue; otherwise**, it will **abort**. |
| Field transformation | COPY | Directly **copies** data from the source field to the target field **without any transformation**. |
| | ADD | **Inserts a new field** into the target schema that does not exist in the source schema. |
| | CAST | **Converts the data type** of the source field to match the expected type of the target field. |
| | DELETE | **Removes the field** from the source schema when it is not required in the target schema. |
| | RENAME | **Changes the name** of the source field to match the name of the target schema. |
| | DEFAULT | **Assigns** a predefined **default value** to a target field when data is unavailable or null. |
| | MISSING | Used when **no appropriate mapping exists** to map the source field to a target field, implying a schema mapping failure. |
| Value transformation | SCALE | **Adjusts the numerical values** in the source field **by a specified factor** according to the value in the target field. |
| | SHIFT | Modifies the values in the source field by **adding or subtracting a constant value**. |
| | LINK | **Establishes a correspondence between values** in the source field and defined values in the target field, used for fields with **enum type**. |
| | GEN | **Generate a transformation function** that **defines how to convert values** from the source field to fit the target field's requirements. |
| | APPLY | **Applies a transformation function**, either generated or predefined by the developer, to the value of a source field to derive the value of the target field. |

```
{from: triggered, to: motion, transformation: RENAME triggered TO motion}
{from: battery_percentage, to: None, transformation: DELETE battery_percentage}
{from: None, to: sensitivity, transformation: ADD sensitivity TYPE integer}
{from: sensitivity, to: sensitivity, transformation: DEFAULT sensitivity TO 2}
{from: enabled, to: enabled, transformation: COPY}
```

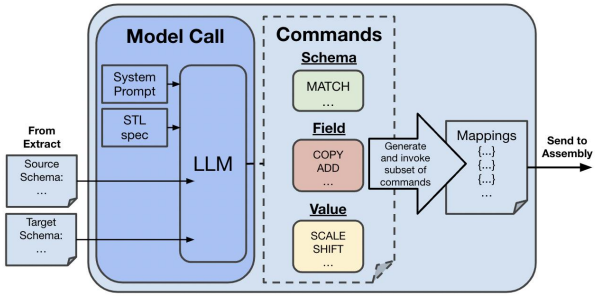
| Source schema | Target schema | Precision | | Recall | | F1 | |
|---|---|---|---|---|---|---|---|
| | | STL | Base | STL | Base | STL | Base |
| Philips Hue | Vivint | **0.91** | 0.73 | **0.98** | 0.83 | **0.94** | 0.78 |
| SimpliSafe | Vivint | **1** | 0.2 | **0.8** | 0.2 | **0.89** | 0.2 |
| SimpliSafe | Philips Hue | **1** | 0.8 | **0.9** | 0.67 | **0.95** | 0.72 |

- Baseline: single model call with a high-level prompt
- **~20%** to **70%** higher mapping accuracy (20 runs, mapping granularity)

**Task: Schema Evolution.** Enable apps/data consumers to automatically adapt to schema changes by data producers.

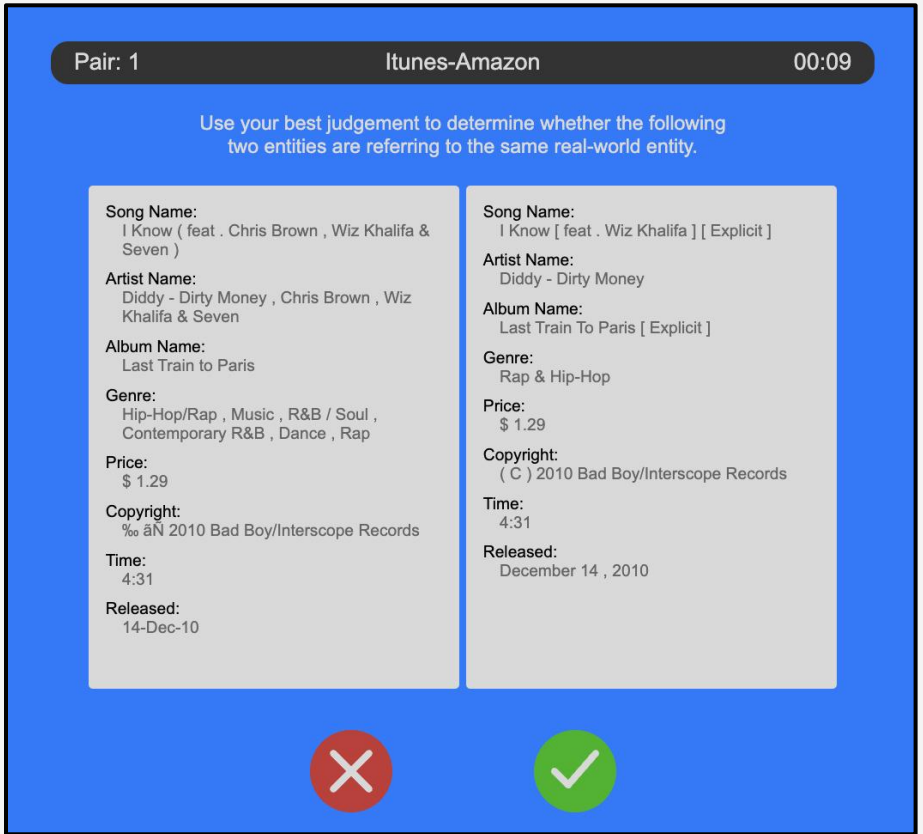
**Today:** simple, limited rule-based automation.

**Approach:** using LLMs to generate schema mappings with a *task-specific language* as an IR, which is then compiled into dataflow ops.



# Compound Schema Registry

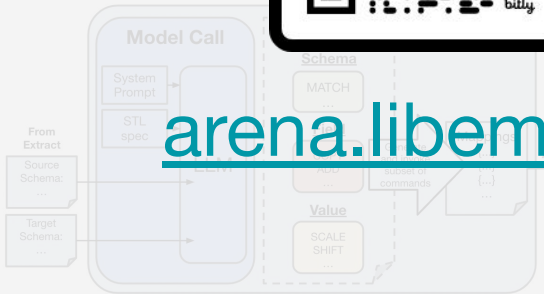
Silvery Fu, Xuewei (Sylvia) Chen
UC Berkeley, System Design Studio




**Task: Schema Evolution.** Enable apps/data consumers to automatically adapt to schema changes by data producers.

**Today:** simple, ed automation.

**Approach:** usi rate schema mappings with anguage as an IR, which is the dataflow ops.




arena.libem.org