Liberal Entity Matching as a Compound AI Toolchain

Silvery Fu[1,2], David Wang[1], Kathleen Ge[1], Wen Zhang[1]
[1]UC Berkeley, [2]System Design Studio

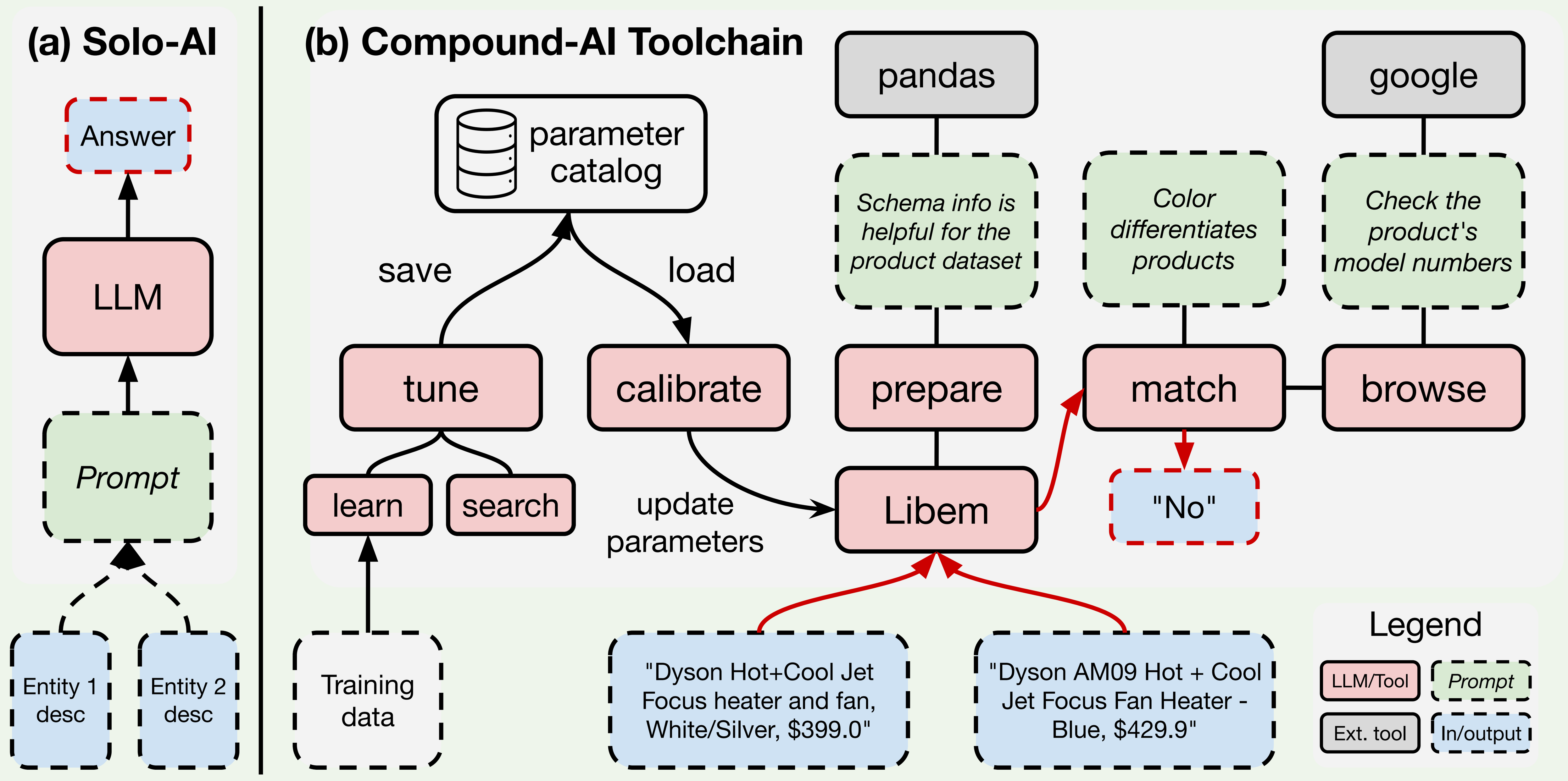
Task: Entity Matching

- Determine whether two **descriptions** refer to the **same entity**, e.g., whether product descriptions on Ebay and Amazon refer to the same product.
- EM solution has evolved from rule-based, crowdsourcing, to deep-learning, PLMs, and recently **LLMs**, showing SoTA results.

Today: Solo-AI EM

- Perform EM in a single model call
- **Challenge:** Hand-tuned prompts, static knowledge, and rigid data preprocessing
- **Proposal:** Compound-AI EM with specialized tools tools and optimizations

Libem: Toolchain for Entity Matching



Liberal Entity Matching with Compound System Designs

Specification

- Decompose EM as sub-tasks and provide each with a specialized tool.
- A model call *liberally* decides what tools to use during EM

Optimization

- Separate parameters and prompts from tools to allow the tools to be optimized using training data and dynamically configured

Composition

- Each tool outputs a confidence level and explanation (via chain of thought), for downstream use of the match result

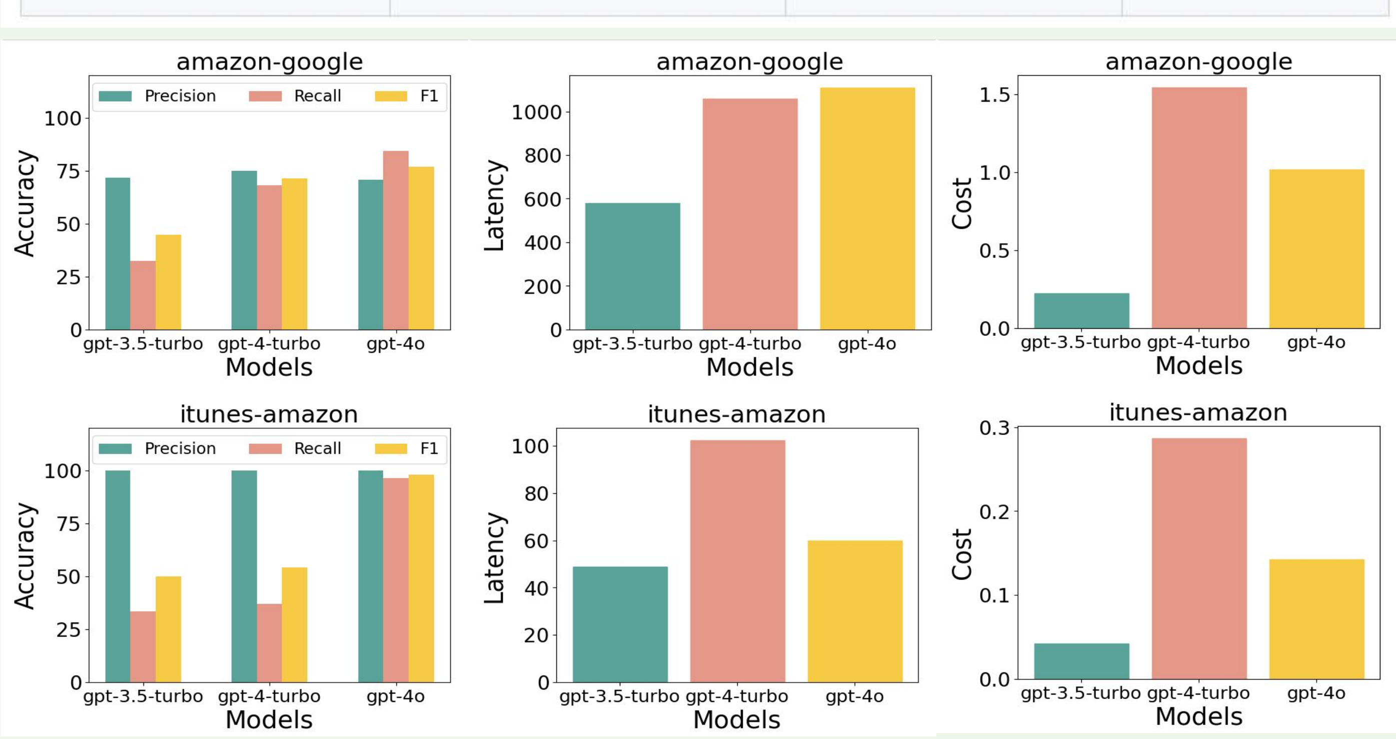
Evaluation

- Solicit human evaluation and labels by gamifying the EM task, and compare Libem performance with human performance.

Solo-AI vs. Libem on EM Datasets

- **Setup: S1:** Solo-AI, **S2:** Compound-AI (Libem); gpt-4o

Dataset	Precision (S1, S2)	Recall (S1, S2)	F1 (S1, S2)
abt-buy	84.0, **89.9**	99.5, 99.5	91.1, **94.5**
amazon-google	60.0, **67.4**	89.7, **92.7**	71.9, **78.1**
beer	92.3, 92.3	85.7, 85.7	88.9, 88.9
dblp-acm	80.4, **94.7**	**100.0**, 99.6	89.1, **97.1**
dblp-scholar	78.4, **88.3**	98.8, **93.6**	87.4, **90.9**
fodors-zagats	95.7, **100.0**	100.0, 100.0	97.8, **100.0**
itunes-amazon	89.3, **100.0**	92.6, **96.3**	90.9, **98.11**
walmart-amazon	75.4, **85.4**	95.3, **91.2**	84.2, **88.2**



- **4.3%** increase in the average **F1** score across the eight datasets, with a maximum of **8.1%** in the itunes-amazon dataset.
- Libem achieved comparable or better performance **without manually tuning** the prompts per-dataset
- **Tool use**, e.g., enabling or disabling schema with *libem.prepare* tool, can substantially **improve matching accuracy**
- For model calls, GPT-4o performs consistently better in terms of **accuracy,** cost, and match latency than 4-turbo and 3.5-turbo

Data Preparation (libem.prepare)

Data Records without Schema (itunes-amazon)

Illusion (feat . Echosmith) Zedd True Colors Dance , Music , Electronic $ 1.29 2015 Interscope Records 6:30 18-May-15

Data Records With Schema

```
"song_name": "Illusion ( feat . Echosmith )",
"artist_name": "Zedd",
"album_name": "True Colors",
"genre": "Dance , Music , Electronic",
"price": "$ 1.29",
"copyright": "2015 Interscope Records",
"time": "6:30", "released": "18-May-15"
```

Browse (libem.browse)

$ libem "mighty strike freedom gundam" "ZGMF/A-262PD-P" → **No**

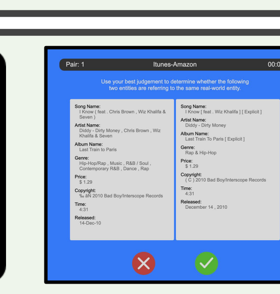
$ libem "mighty strike freedom gundam" "ZGMF/A-262PD-P" **--browse**
[browse] search duckduckgo: ZGMF/A-262PD-P
[browse] ..pilots the Mighty Strike Freedom Gundam. Prototype figure..
Match: **Yes**

Confidence and Explanation

$ libem "dyson fan+heater am09" "dyson hp purifier" **--confidence --cot**
Explanation: 1. **Brand**: Both entities are Dyson product
2. **Product Type**: "fan+heater" vs. "purifier" …
Match: no; **Confidence: 5**

Arena

- Libem Area is a crowdsourcing app to "gamify" the data labeling process
- Test your matching skills against other users and Libem in this demo right now →



arena.libem.org

Project: https://github.com/abcsys/libem/