ONE THING, MANY NAMES

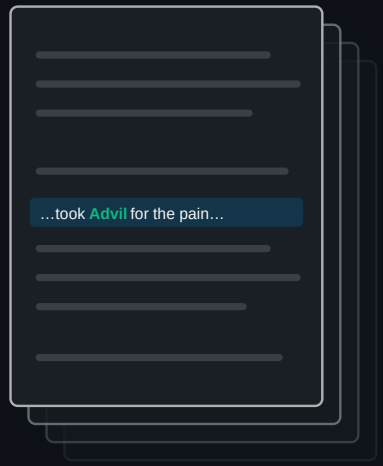
Building Reliable Entity Mapping

From LLM extraction to reliable, queryable knowledge



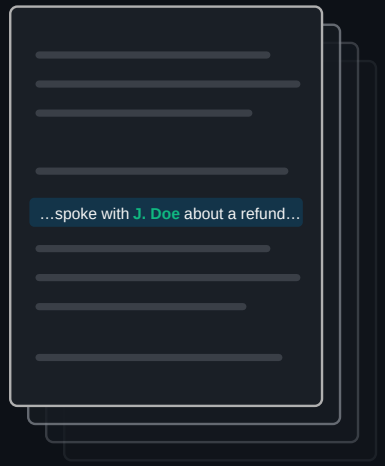
Maryam Astero | Genomenon | LLMday Austin | October 14, 2026

LLMs made extraction scalable



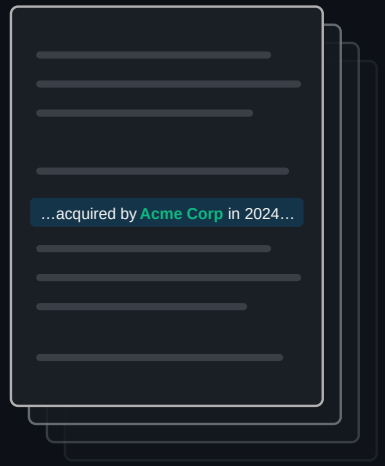
scientific papers

a drug



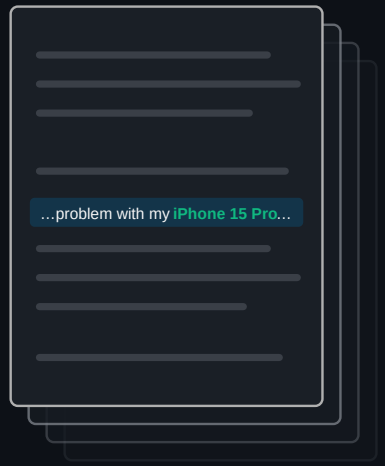
support tickets

a customer



financial filings

an organization



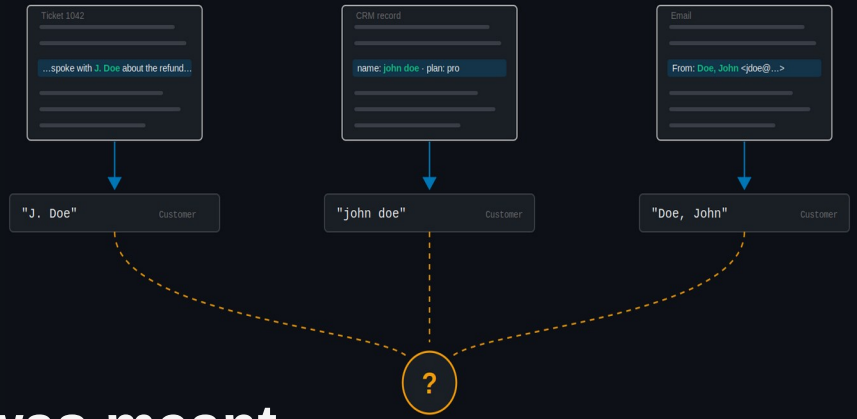
product reviews

a product

But what did we actually produce?

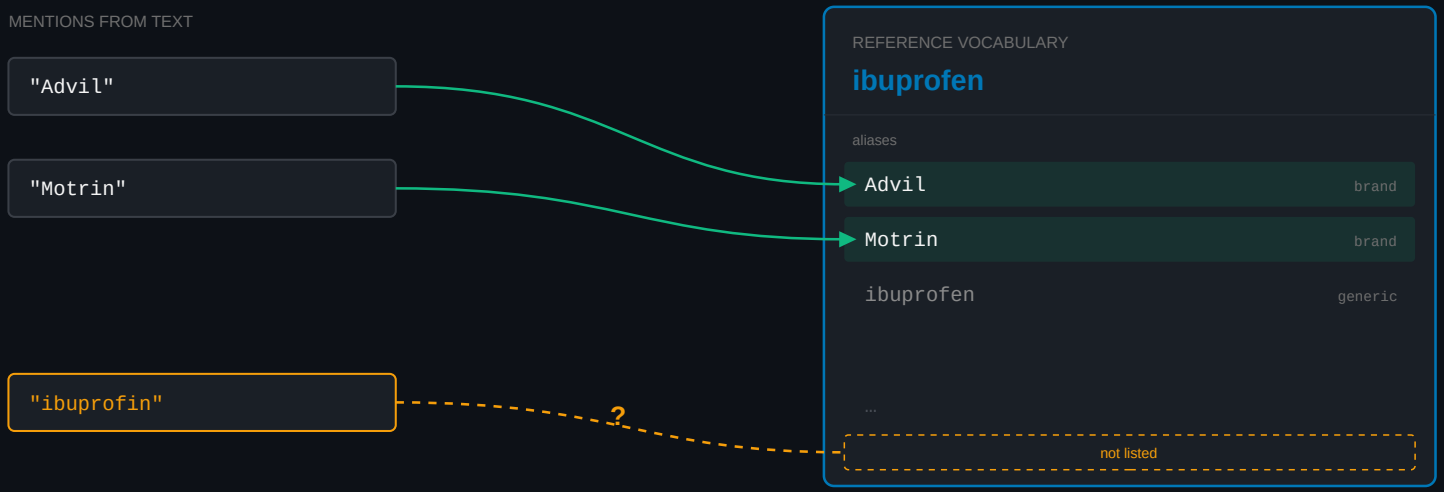
A mention is not an entity!





Extraction = what was written
Mapping = which real-world entity was meant

Can the reference data solve it?



What happens when the exact name isn't in the reference data?

Can we find a close name anyway?

mention	i	b	u	p	r	o	f	i	n
reference	i	b	u	p	r	o	f	e	n

one wrong character
→ **candidate found**

---

mention	s	w	o	l	l	e	n		a	n	k	l	e	s
reference	a	n	k	l	e		s	w	e	l	l	i	n	g

same words, different order and form
→ **same concept, no candidate**

Can we find it by meaning?

Can we find it by meaning?

MENTION

`"swollen ankles"`

not in the vocabulary · no string tier matched
shares no word with the canonical name

semantic retrieval

CANDIDATES, distinct entities, ranked by similarity

1	**Ankle swelling**	the concept the mention describes	**highest semantic similarity → top candidate**
2	**Enlargement of the ankles**	a related but different finding	
3	**Ankle joint effusion**	fluid in the joint, a different finding	

no string method could see it. The embedding ranked it first, with a wide margin over the runner-up.

So is the top candidate the answer?

Is the top candidate the identity?

MENTION

"Advil PM"

contains the word Advil

CANDIDATES, ranked by similarity

1 **Advil** — ibuprofen

2 **Genexol PM** — a chemotherapy drug

3 **Refresh PM** — an eye ointment

DECISION

?

Which one do we map to?

Is the top candidate the identity?

MENTION

"Advil PM"

contains the word Advil

CANDIDATES, ranked by similarity

1	**Advil**	ibuprofen	**similar. not the same product.**
2	**Genexol PM**	a chemotherapy drug	**not Advil PM**
3	**Refresh PM**	an eye ointment	**not Advil PM**
	Advil PM · ibuprofen + diphenhydramine		no canonical entity found in the vocabulary

DECISION

TOP-1 RULE → ibuprofen

The best available match was Advil. The correct product was not in the vocabulary.

Do we have enough evidence to map this mention to any of them?

What evidence do we actually have?

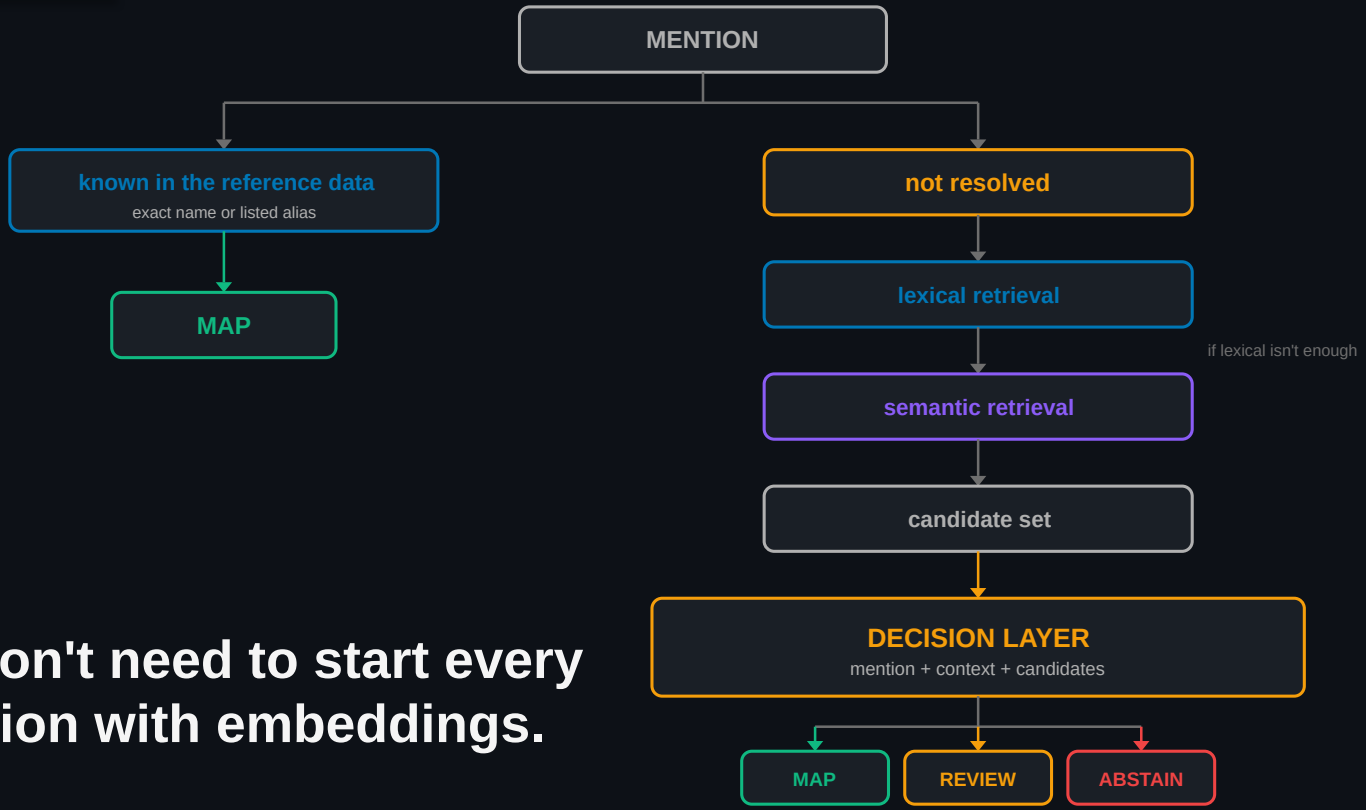
Exact / alias	→ The reference data explicitly links this name to the entity.
Lexical	→ These strings are very similar.
Semantic	→ These concepts are related or plausible.
Context	→ This entity fits what the document is saying.
Competition	→ This candidate is clearly separated from the alternatives.

The question isn't: which model should we use?

The question is: is the evidence sufficient for this decision?

Similarity is evidence. It is not identity.

Use the simplest evidence that is sufficient



We don't need to start every mention with embeddings.

What if the evidence isn't sufficient?

Strong evidence	→	**MAP**
Conflicting evidence	→	**REVIEW**
Insufficient evidence	→	**ABSTAIN**

A wrong connection propagates as if it were a fact.

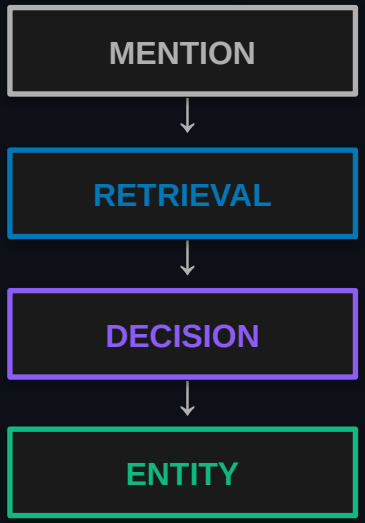
If we map anyway:

WRONG MAPPING
↓ graph / search / analytics / agents
↓ wrong downstream knowledge

Not mapping is sometimes the correct decision.

So where, exactly, is the system failing?

How do we know what's failing?



Did the right entity enter the candidate set?

Did we choose correctly?

AMBIGUITY

Is there enough information in the text?

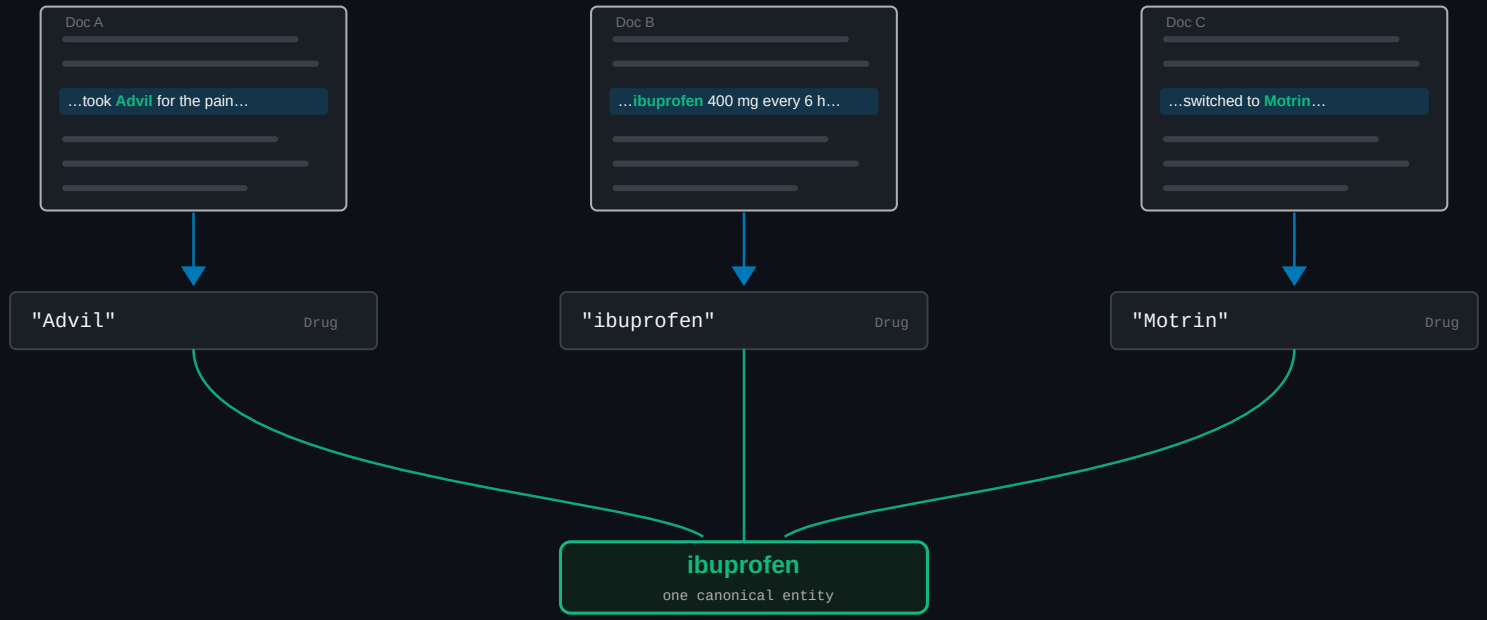
REFERENCE DATA

Does our vocabulary contain the entity or alias?

Different failures, different fixes. A better model may help with one of them.

One end-to-end number hides where the problem is.

From mentions to connected knowledge



Extraction gives you mentions. Mapping makes them connectable.

Thank you. Questions?

Let's connect