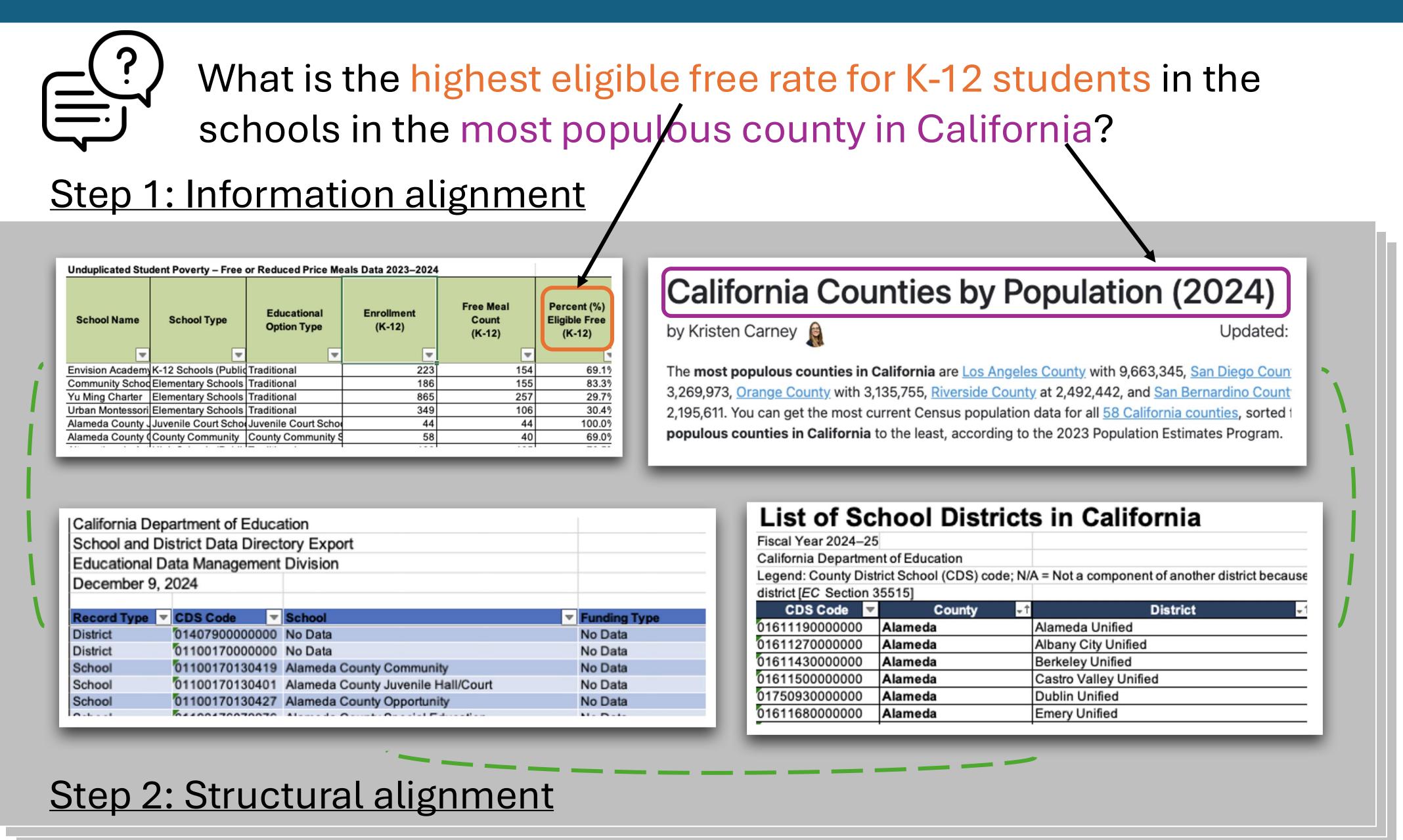# Can we Retrieve Everything All at Once?
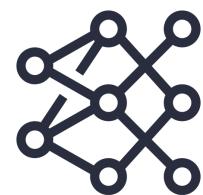# ARM: An *A*lignment-Oriented LLM-based *R*etrieval *M*ethod
Peter Baile Chen (MIT), Yi Zhang (AWS AI), Michael Cafarella (MIT), Dan Roth (UPenn)

SCAN ME

An optimal retrieval system should consider both *information alignment* and *structural alignment*

What is the highest eligible free rate for K-12 students in the schools in the most populous county in California?

Step 1: Information alignment

Step 2: Structural alignment

Step 3: Self-verification and aggregation

Vote for each object based on
- Relevance between user query and object content
- Compatibility within objects

Return the top-k objects with the highest votes

End-to-end performance

	Bird	OTT-QA	2WikiMultiHop
Model for answer generation: Llama-3.1-8B-Instruct			
Standard RAG	17.5	39.8	31.7
Agentic RAG	4.7	26.0	33.8
ARM$_{Llama-3.1-8B}$	**20.6**	**44.0**	**45.0**
Model for answer generation: GPT-4o-mini			
Standard RAG	29.8	45.9	41.7
Agentic RAG	26.3	41.2	52.9
ARM$_{Llama-3.1-8B}$	**31.8**	**49.1**	**53.6**

Monetary cost

Cost of Agentic RAG relative to ARM	Llama-3.1-8B-Instruct	GPT-4o-mini
Bird	5.36x more expensive	3.07x more expensive
OTT-QA	4.07x more expensive	3.38x more expensive
2WikiMultiHop	1.44x more expensive	1.31x more expensive