Revisiting Composite Item Retrieval in the Era of Large Language Models: Challenges and Opportunities

Senjuti Basu Roy

senjutib@njit.edu

Senjuti Basu Roy

Panasonic Endowed Chair of Sustainability

Associate Professor, Dept of Computer Sc. (https://web.njit.edu/~senjutib/)

Director: Big Data Analytics Laboratory

https://centers.njit.edu/bdal

Research Profile and Interests

- Human-in-the-loop AI,
- Future of Work
- Responsible Data management
- Over $11M of research awards
- 90+ publications, ≥ 4000 citations, h-index ≥ 30

Current services

- Associate Editor ACM Very Large Database Journal
- Associate Editor ACM International Conference of Management of Data
- Associate Editor of Information Systems
- National Science Foundation Panel

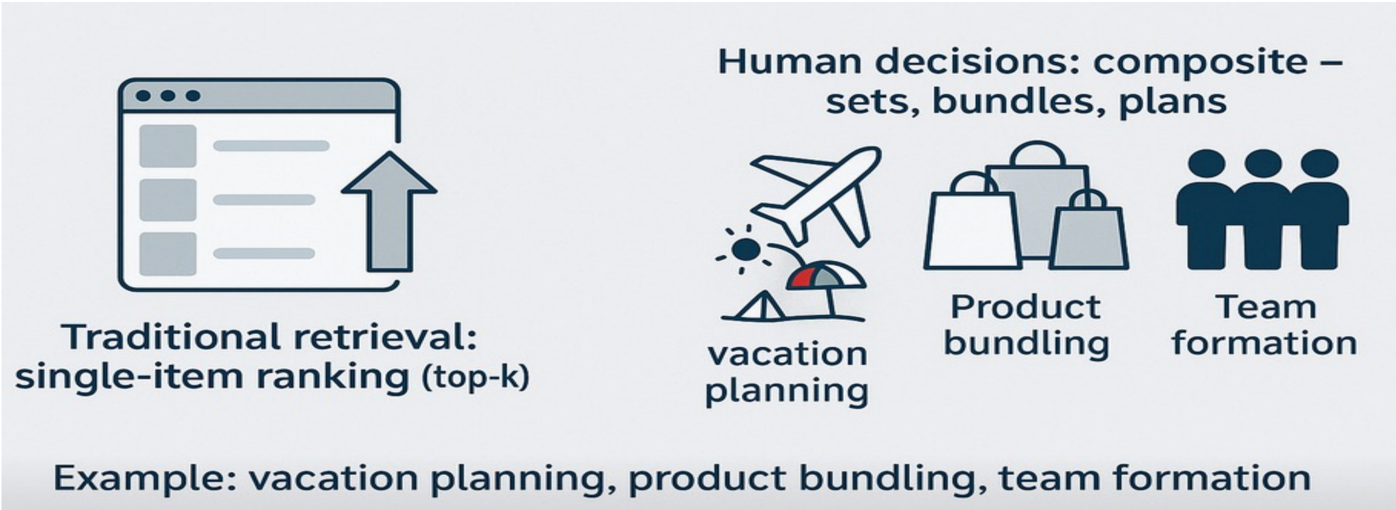
Awards and honors

- NSF Career Award 2020
- Presidential Early Career Awards for Scientists and Engineers (PECASE) Nominee
- Recognized as one of the 100 early career engineers by the National Academy of Engineers(NAE)
- An Inventor Member of the National Academy of Inventors (NAI
- Acknowledged and Profiled as one of the three female researchers by Microsoft Research in Grace Hopper Computing in 2014

Overview

- Part I – Context & Foundations
- Part II – LLM Disruption
- Part III – Opportunities & Challenges
- Part IV – Vision - From Retrieval to Responsible Reasoning

From Items to Composition : How do we retrieve sets that make sense together?



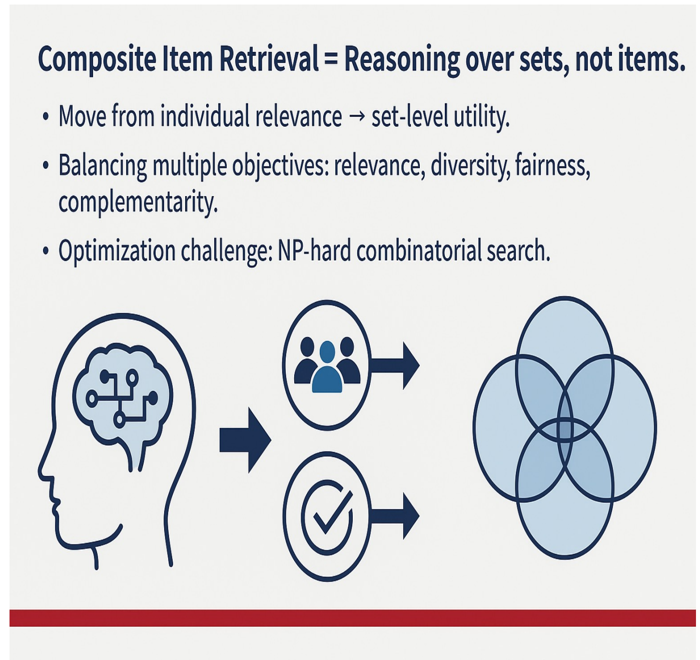
- Traditional retrieval: single-item ranking (top-k).
- Human decisions: composite — sets, bundles, plans.
- Example: vacation planning, product bundling, team formation.

Composite Retrieval

Big Data Analytics Lab

Composite retrieval is the study of methods for ***building, retrieving and ranking composite responses from a set of atomic ones***

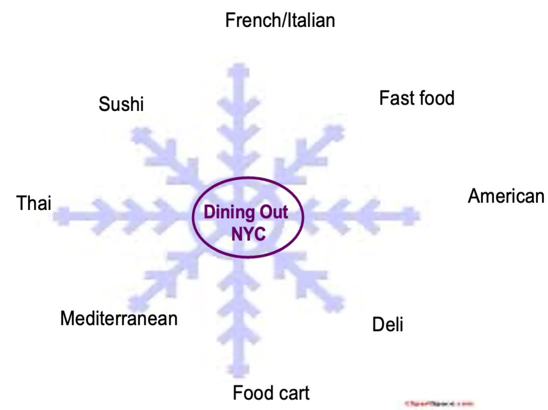
- ❑ Online shopping
- ❑ Web search
- ❑ Recommendation



1.Basu Roy, Senjuti, et al. "Constructing and exploring composite items." *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 2010.
2. Amer-Yahia, Sihem, and Senjuti Basu Roy. "Interactive exploration of composite items." *International Conference on Extending Database Technology (EDBT)*. 2018.
3. Amer-Yahia, Sihem, and Senjuti Basu Roy. "From Complex Object Exploration to Complex Crowdsourcing." *Proceedings of the 24th International Conference on World Wide Web*. 2015.
4. Roy, S. B., Das, G., Amer-Yahia, S., & Yu, C. (2011, April). Interactive itinerary planning. In *2011 IEEE 27th International Conference on Data Engineering* (pp. 15-26). IEEE.
5. Roy, S. B. (2019, January). Human-in-the-loop Exploration of Composite Items. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data* (pp. 367-367).

Composite Item Retrieval

Star Retrieval



Snowflake Retrieval

Chain Retrieval

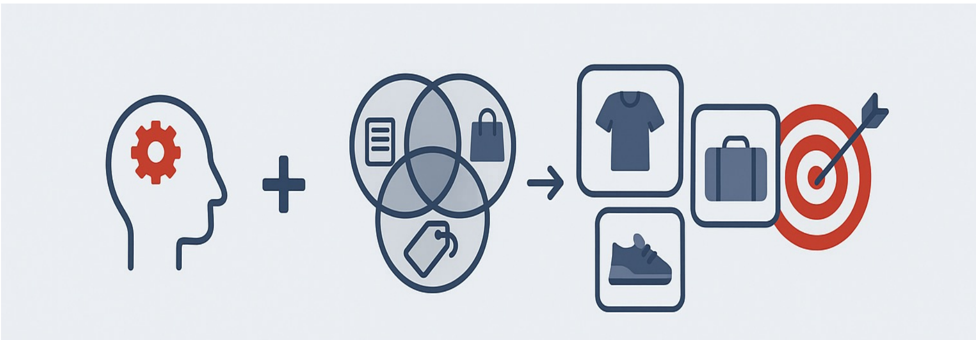
Properties of Composite Item Retrieval

- **Expressed as a single-objective optimization problem with constraints**
- **Input:** a set of items, constraints
- **Output**: k-Composite Items (CIs)
- **Constraints**
  - *Compatibility* between items forming a CI (usually pairwise)
  - *Validity***:** total cost of items forming a CI (e.g., price, time)
  - *Size*: in terms of number of items forming a CI
  - *Type coverage*: multiplicity bounds on item types in a CI
- **Objective function**
  - *Coverage of items*
  - *Diversity* of CIs
  - Additive/Coverage-based/Complex

The Problems and The Algorithms

Problem Name	CI Shape	Hardness	Reduction
Composite k-Package Recommendation (RecSys 2010)	star	NP-hard	Remains hard, even for k=1, reduction from the Knapsack problem
Chain retrieval (ICDE 2011)	chain	NP-hard	Rooted Orienteering problem
KOR Query (VLDB 2012)	chain	NP-hard	Weight Constrained Shortest Path Problem
Diverse k-composite Package retrieval (TKDE 2014)	snowflake	NP-hard	Maximum Edge Subgraph Problem
TourRec: Additive Tour (WSDM 2014)	chain	NP-hard	Traveling Salesman Problem (TSP)
TourRec: CoveringTour	snowflake	NP-hard	Maximum-k Coverage Problem
Star retrieval (SIGMOD 2010) a. maximal package retrieval b. Summarization c. Diversified ordering	star	#P-Complete NP-hard NP-hard	Requires solving the Counting Problem Reduction from the Set Cover Reduction from the TSP

To Summarize – the Core Idea



The Core Idea

Composite Item Retrieval = Reasoning over sets, not items.

- Move from individual relevance → set-level utility
- Balancing multiple objectives: relevance, diversity, fairness, complementarity
- Optimization challenge: NP-hard combinatorial search

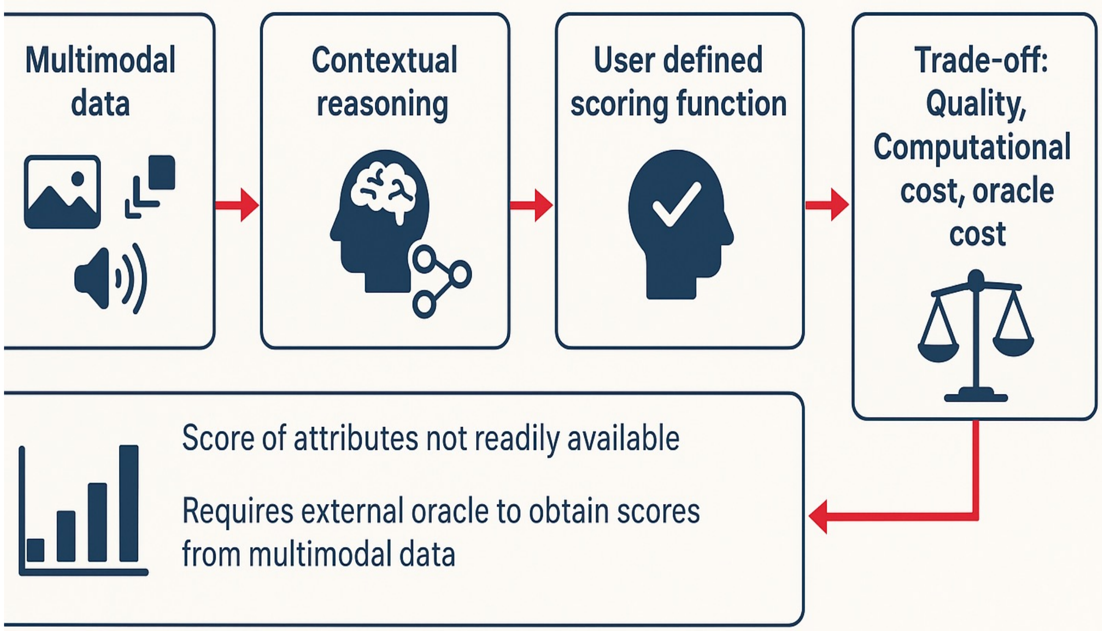
Foundations: submodular optimization, skyline queries, top-k aggregation.

Earlier focus: efficiency, scalability, static user intent.

Assumptions: structured data, clear scoring functions.

PART II: LLM Disruptions

LLM Disruptions



- Multimodal data
  - 80-90% of enterprise data is unstructured
- Personalization
  - Implicit understanding of user intent
  - User defined scoring function
- Contextual reasoning
  - Bring context in understanding user intent
- Score of attributes not readily available
  - Requires external oracle

Trade-0ff: Quality, Computational cost, oracle cost

LLM Disruption — What Changes

Traditional Retrieval	LLM-Era Retrieval
Explicit query and scoring function	Implicit, learned understanding of user intent
Deterministic data and fixed schema	Probabilistic, context-dependent reasoning
Known, static utility function	Dynamic, user-adaptive utility inferred from dialogue
Symbolic matching and optimization	Neural inference and contextual grounding
Independent item ranking	Joint reasoning across multimodal, interdependent items
Limited to retrieval	Extends to explanation, synthesis, and decision-making

PART III –Opportunities & Challenges

Opportunities and Challenges with LLMs

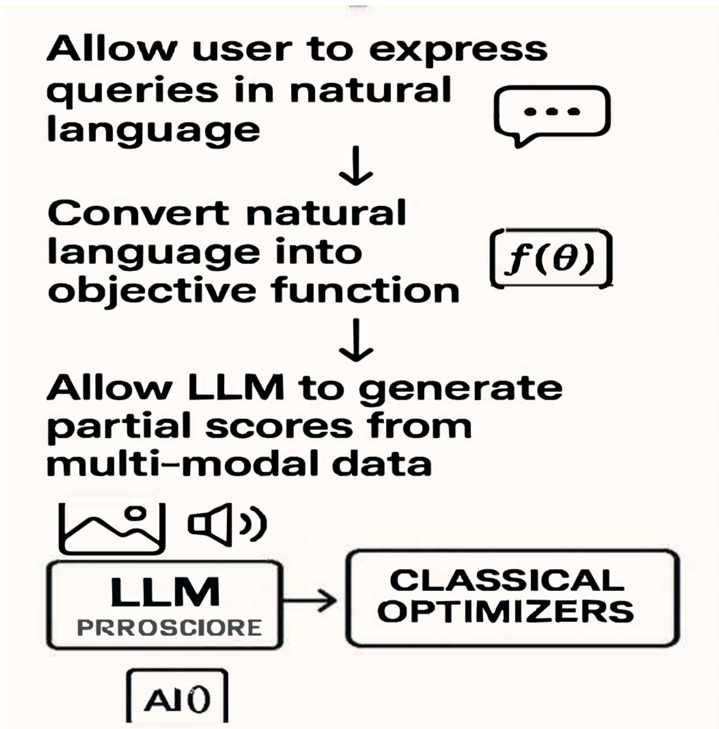
- **Opportunities**
  - **Personalization:** User provides intent to guide retrieval and reasoning.
  - **Contextual Reasoning:** Enables understanding of how items relate logically or semantically.
  - **Leveraging LLMs:** Large Language Models can be used to score subgoals and infer missing relationships.

- **Challenges:**
  - **Query Decomposition:** Breaks complex queries into subgoals with implicit understanding of dependencies.
  - Risk of *hallucination*, leading to uncertain reasoning and unreliable scores.

Computational and inference costs associated with LLM usage - Trade-off between accuracy, efficiency, and cost!!

Hybrid Architectures

- Allow user to express queries in natural language
- Convert natural language into objective function using AI models
- Judiciously leverage LLMs for score generation
- Use classical optimizers for scoring.



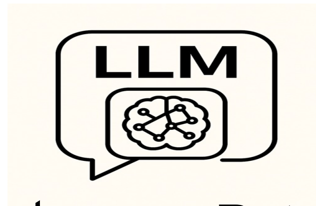
Nia, Sohrab Namazi, Subhodeep Ghosh, Senjuti Basu Roy, and Sihem Amer-Yahia. "Personalized Top-k Set Queries Over Predicted Scores." *arXiv preprint arXiv:2502.12998* (2025).

Nikookar, S., Namazi Nia, S., Basu Roy, S., Amer-Yahia, S., & Omidvar-Tehrani, B. (2025). Model reusability in Reinforcement Learning. *The VLDB Journal*, *34*(4), 41

Large Language Models Empowered Personalized Web Agents (WWW'25) — introduces *LLM-based personalized web agents* capable of adaptive, context-aware recommendations through natural language reasoning

Proposed Framework

Big Data Analytics Lab



Judiciously involve LLM to return score from multimodal data

Return score (handle uncertainty)

Understand user intent

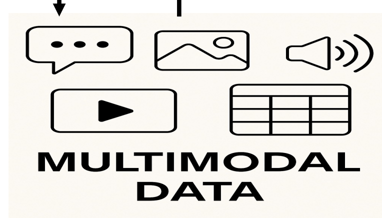
A computational loop leveraging LLM

Perform combinatorial optimization

Output

"Show me top-3 sustainable cocktail attire that are Stylish, not overly formal; flattering but not flashy."

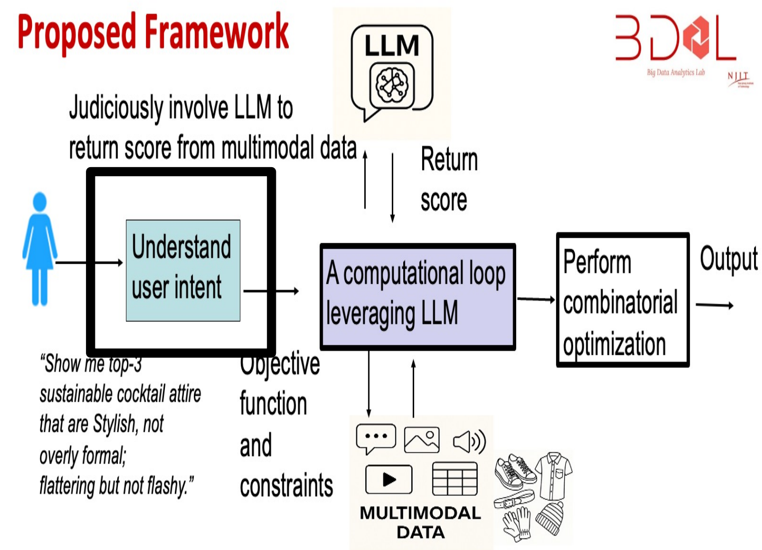
Objective function and constraints



Nia, Sohrab Namazi, Subhodeep Ghosh, Senjuti Basu Roy, and Sihem Amer-Yahia. "Personalized Top-k Set Queries Over Predicted Scores." *arXiv preprint* (2025).
Nikookar, S., Namazi Nia, S., Basu Roy, S., Amer-Yahia, S., & Omidvar-Tehrani, B. (2025). Model reusability in Reinforcement Learning. *The VLDB Journal*, *34*(4), 41

A Reinforcement Learning Based Framework — Understanding User Intent

Convert user intent into interpretable **components**:



- **Goal type** → optimization, retrieval, recommendation, classification, reasoning. *"sustainable cocktail attire that are Stylish, not overly formal; flattering but not flashy"* → multi-objective optimization.
- **Objective dimensions** → relevance, diversity, serendipity, etc.
- **Constraint type** → hard constraints ("must not exceed certain price") vs soft preferences ("prefer certain colors").

$$Intent = < Goal\ Type, Objectives, Constraints, Weights >$$

$$A \longrightarrow B \longrightarrow C \longrightarrow D$$

).

Nikookar, S., Namazi Nia, S., Basu Roy, S., Amer-Yahia, S., & Omidvar-Tehrani, B. (2025). Model reusability in Reinforcement Learning. *The VLDB Journal*, *34*(4), 41

Leverage LLM in the Computational Loop

Big Data Analytics Lab

NJIT

- **Task : Retrieve relevance and diversity score of fashion brands**
- ***Scoring function***

$$F(s,u) = \sum_{e \in s} Rel(u,e) + \sum_{e_i, e_j \in s} Div(e_i, e_j)$$

Entity ID	Name	Items
1	VEJ	$\{i_1, i_3\}$
2	REF	$\{i_2, i_5\}$
3	MAT	$\{i_6, i_9\}$
4	BOO	$\{i_4, i_8\}$
5	EVR	$\{i_7, i_{10}\}$

Fashion brands

Brand Name	Rel
Veja (VEJ)	U
Reformation (REF)	1.0
Matt & Nat (MAT)	1.0
Boohoo (BOO)	0.0
Everlane (EVR)	0.5

Relevance

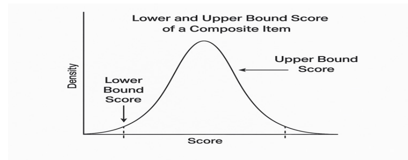
	VEJ	REF	MAT	BOO	EVR
VEJ	-	1.0	0.5	0.5	0.5
REF	1.0	-	U	U	U
MAT	0.5	U	-	U	0.5
BOO	0.5	U	U	-	U
EVR	0.5	U	0.5	U	-

Diversity

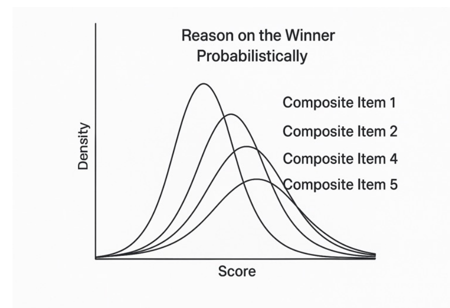
How can we select the most informative partial scores from the LLM to maximize accuracy and minimize latency?

4 Computational Steps

Reason on the score of the candidates
- Reason on score bounds as opposed to full score



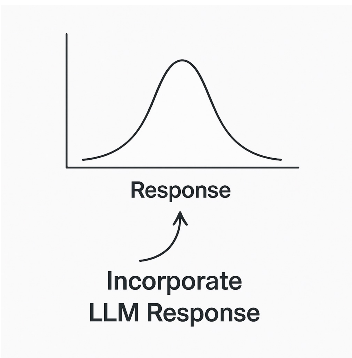
Reason on the winner probabilistically
- What is the probability of a composite item to be the top-1 set based on the objective function



Identify the next best question to be asked to LLM

which component of the unknown score for which items queried from the LLM next

Incorporate LLM response in the loop
- LLM can hallucinate and provide non-deterministic response



Nia, Sohrab Namazi, Subhodeep Ghosh, Senjuti Basu Roy, and Sihem Amer-Yahia. "Personalized Top-k Set Queries Over Predicted Scores." *arXiv preprint arXiv:2502.12998* (2025).

Algorithmic Framework

Repeat Until Stopping condition is reached

- Discrete LLM Response -> Stop when only one single candidate is left
- Range LLM response ->Stops when confidence ≥ threshold θ

1. **Compute score bounds (LB, UB) of candidate CIs ->** Prune candidates based on bounds
2. **Decide winner probabilistically**

$$P(c = c^*) = \prod_{i=1}^{M-1} P\left(\mathcal{F}(c,u) \geq \mathcal{F}(c_i,u) \mid \bigcap_{j=1}^{i-1} (\mathcal{F}(c,u) \geq \mathcal{F}(c_j,u)) \right). \quad (3)$$

$$P(c = c^*) = \prod_{i=1}^{M} P(\mathcal{F}(c,u) \geq \mathcal{F}(c_i,u)).$$

3. **Ask Next Question to LLM based on uncertainty reasoning**

$$H(c^*) = -\sum_{i=1}^{n} p(c_i = c^*) \log p(c_i = c^*)$$

4. **Process LLM response back in the loop**

Scalability Challenges in Probabilistic Modeling for Finding Winner

Candidates are Independent

- Independence model: uses convolution. Time Complexity: $\Theta(M^2 \times m)$
- Memory: Linear in m
- Strengths: Efficient, lightweight, scalable for large candidate sets.
- Challenges: Independence assumption does not hold when candidates overlap

Candidates are Dependent

- Models shared entities using conditional probabilities.
- Optimization: Computes probabilities pairwise, avoiding $O(m^M)$ explosion.
- Time Complexity: $\Theta(M^2 \times m^2)$
- Space Complexity: $O(m^2)$
- Strengths: Accurately models dependencies.
- Challenges: High computation and storage cost, limited parallelizability.

Overall Scalability Challenges

- Quadratic growth with candidate count (M).
- Quadratic convolution cost with discretization size (m).
- Overlapping entities → dependency propagation → cost escalation.
- Maintaining PDFs and bounds adds compute/memory overhead.
- Multi-modal data intensifies scalability constraints.
- Trade-off: ProbInd = efficient but approximate; ProbDep = accurate but expensive.
- Balancing scalability vs. dependency modeling is key.

Experimental Evaluation

- **Datasets**: Hotels, Movies, Yelp Businesses
- **LLMs**: GPT-4o mini, Llama-3-70B
- **Metrics**: #probes, cost, latency, scalability, user study
- **Example scoring functions:** (F1–F7) with Relevance & Diversity pairs
- **Baselines**: Random, Full-Probe

Scoring	Relevance	Diversity
$\mathcal{F}_1$	Hotel rating	Geo distance
$\mathcal{F}_2$	Proximity to city center	Star rating
$\mathcal{F}_3$	Brief plot	Production year
$\mathcal{F}_4$	Popularity	Genres & eras
$\mathcal{F}_5$	Location near New York	Cost variety
$\mathcal{F}_6$	Cuisine type	Operating hours
$\mathcal{F}_7$	Similar to W. Anderson Movies	Diff. decades

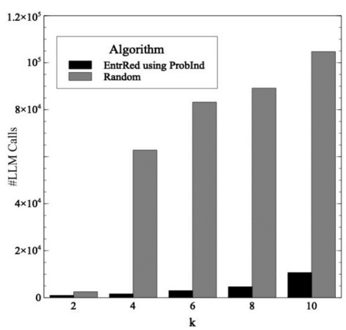
Table 5: Personalized scoring functions used in experiments

Experimental Evaluation

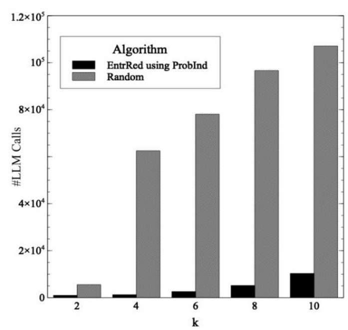
Big Data Analytics Lab

- Key Takeaways:
  1. 10 × fewer LLM probes vs baselines
  2. Same top-k accuracy (100% recall)
  3. Monetary Cost ↓ by order-of-magnitude (e.g., \$14 → \$1)
  4. ProbDep slightly better but slower than ProbInd

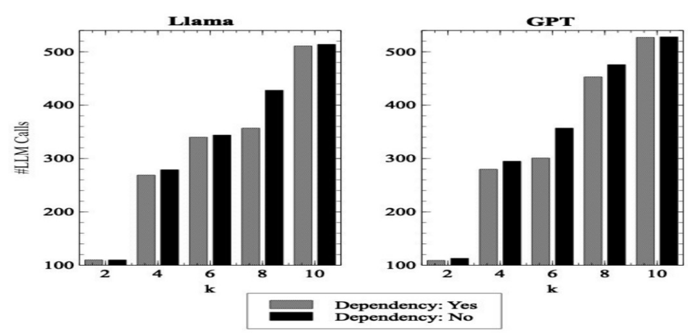
- User study:
  - 80–95% of users preferred our recommendations (Movies dataset)



(a) Hotels - $\mathcal{F}_1$ - Discrete



(b) Businesses - $\mathcal{F}_5$ - Discrete



(a) Hotels - Scoring function $\mathcal{F}_1$

IV: Vision - From Retrieval to Responsible Reasoning

Key Dimensions

- **Ambiguity and Intent Understanding**
  - Translating user queries into well-defined reasoning objectives.
  - Handling underspecified or conflicting goals in natural language.
  - Balancing task completion with ethical interpretation of intent.

- **Multi-Modal Integration**
  - Combining heterogeneous sources (text, image, structured data, sensor signals).
  - Aligning representations across modalities for coherent reasoning.

- **Fairness, Bias, and Explainability**
  - Preventing propagation of bias from retrieved data to reasoning outcomes.
  - Designing explainable reasoning chains that justify conclusions.
  - Incorporating fairness constraints into reasoning pipelines.

- **Robustness and Uncertainty Handling**
  - Propagating uncertainty from retrieval through multi-step reasoning.
  - Avoiding overconfidence in generative or inferential steps.
  - Probabilistic reasoning under incomplete or noisy evidence.

- **Scalability and Efficiency**
  - Moving from single-query retrieval to continuous reasoning over dynamic data streams.
  - Integrating symbolic and neural reasoning efficiently.
  - Maintaining real-time performance under large-scale multi-modal data.

Reasoning with AI in Big Data Analytics Lab

Welcome

Big Data Analytics Lab

The Big Data Analytics Lab (BDaL), is an interdisciplinary research laboratory, that focuses on large-scale data analytics problems that arise in different application domains and disciplines. One of the primary focus of our lab is to investigate an alternative computational paradigm that involves "humans-in-the-loop" for large-scale analytics problems. These problems arise at different stages in a traditional data science pipeline (e.g., data cleaning, query answering, ad-hoc data exploration, or predictive modeling), as well as from emerging applications.

We study optimization opportunities that come across because of this unique man-machine collaboration and address data management and computational challenges to enable large-scale analytics with humans-in-the-loop. Our focus domains are social networks, healthcare, climate science, retail and business, and spatial data. The research projects at BDaL are funded by the National Science Foundation, Office of Naval Research, National Institute of Health, and Microsoft Research.

Project 1: Human AI Agile Symbiosis

Sponsor: Department of Defense

Goal: A framework to enable proactive, context-dependent decision support with enhanced operational capability under uncertainty

Project 2: Predictive Modeling for Ship Scheduling (PASS)

Sponsor: Department of Defense

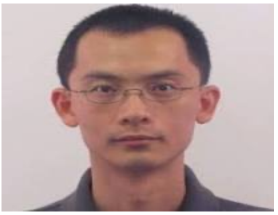
Goal: Human Compatible Decision Support Systems for Planning and Actual Maintenance of US Naval Ships

Project 3: Form Curation While Creation

Sponsor: Department of Defense

Goal: Leverage gen AI to generate forms to write natural language texts that aid sailors

Key Collaborators





PhD students

Sohrab Namazi Nia
Thinh On
Swastik Biswas
Subhodeep Ghosh

Postdocs and Research Scientists

Gerald White
Kevin Chhoa
Deep Mistry
Manish Kumar

3DAL
Big Data Analytics Lab
NJIT

NSF

Office of Naval Research
ONR
Science & Technology

NIH
National Institutes of Health

Microsoft Research

Thank You!