SAN JOSE STATE UNVERSITY

Escape the Big Data Trap: LLMs Liberate Recommenders *via* *Training-Free Condensation*

Harshith Akkapelli

CMPE 255 - Harshith Akkapelli

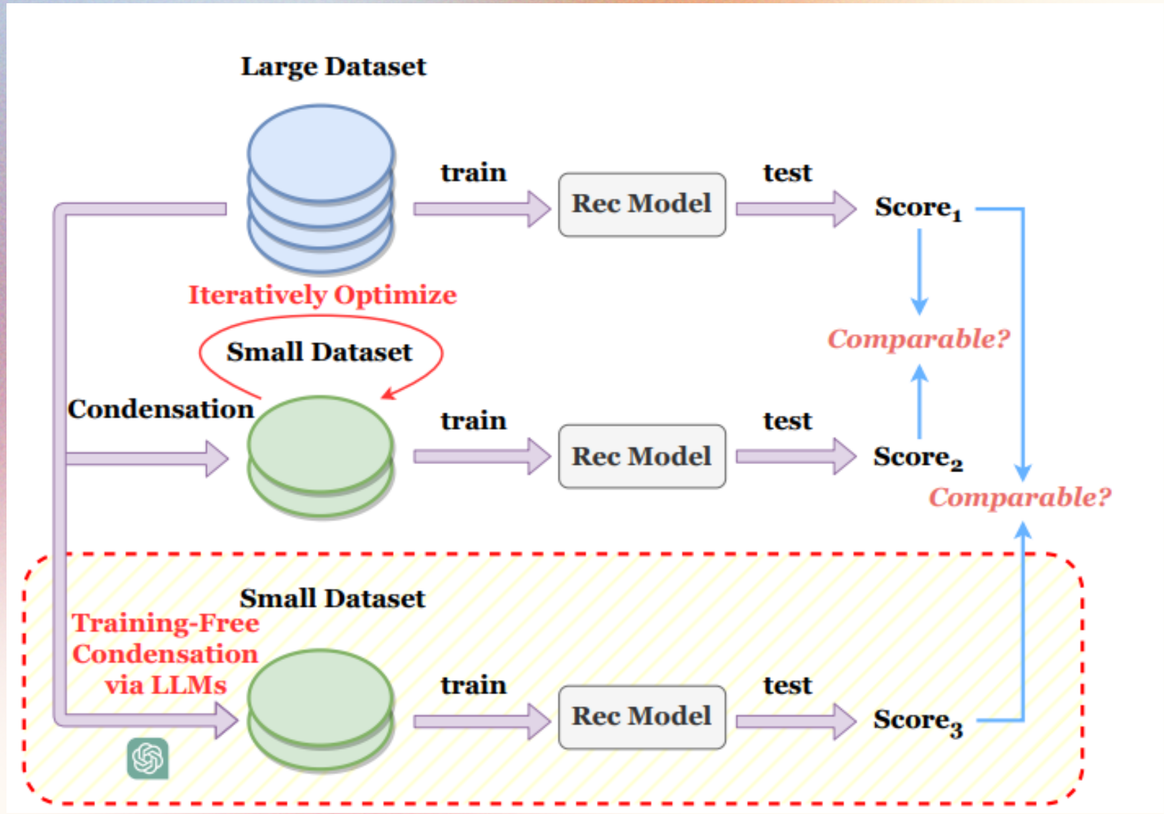
Course Instructor: Vijay Eranti

Why Training Data Condensation?

Recommendation is a crucial part in today's large scale systems which really improves user experience. These days the amount of data is extremely huge, especially text data. Training the recommendation system with huge data is an extremely cumbersome process. Therefore condensing the large dataset to a valuable small subset of data is extremely crucial.

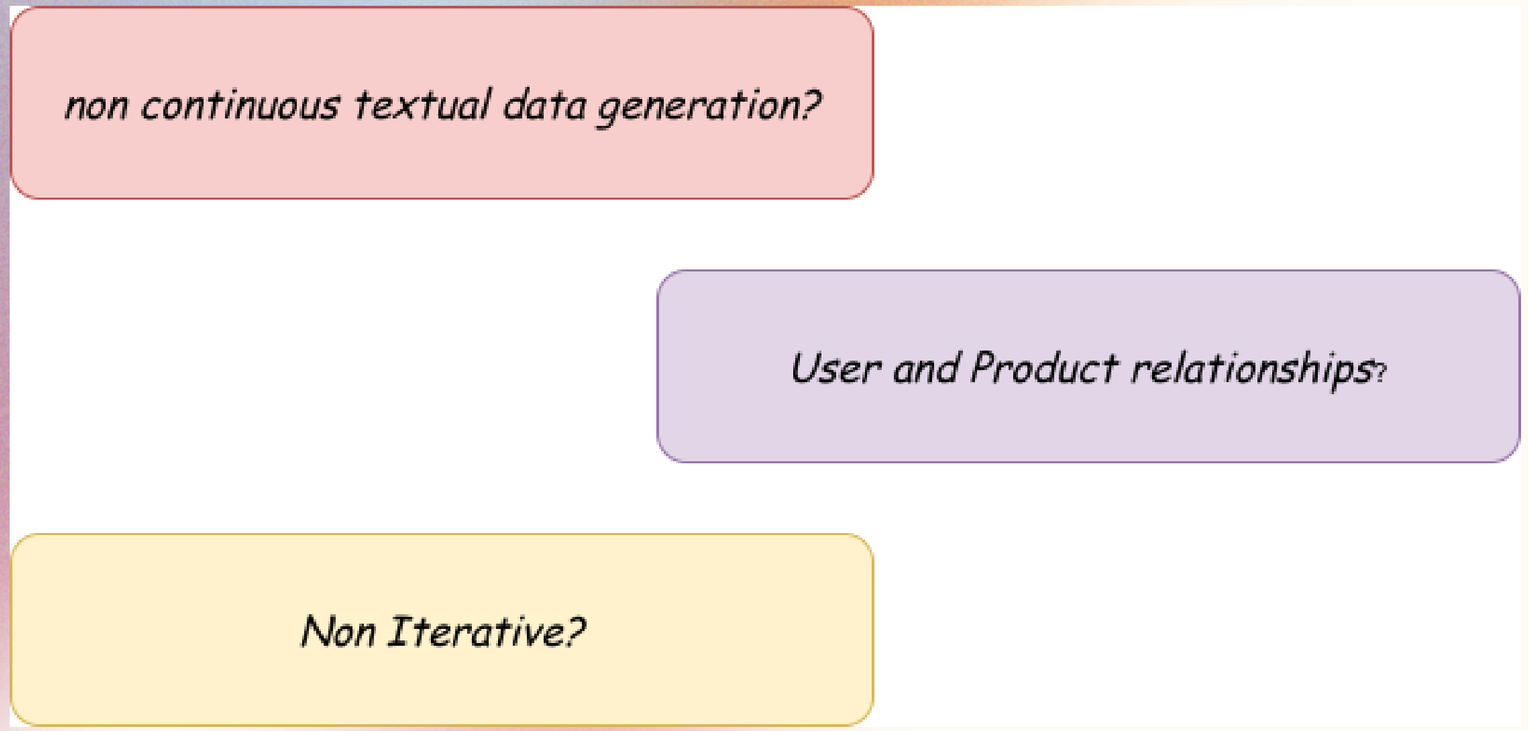
Existing Solutions and their Problems

1. They are iterative meaning the condensed dataset have to be iterated over time to get better performance
2. They do not capture the relationship between users and items
3. They cannot generate new text data which is non continuous



New Approach

1. It is not an iterative approach rather a one way or forward process
2. It is primarily for textual data which is discrete data
3. It makes sure to capture the relationships between user and items/products



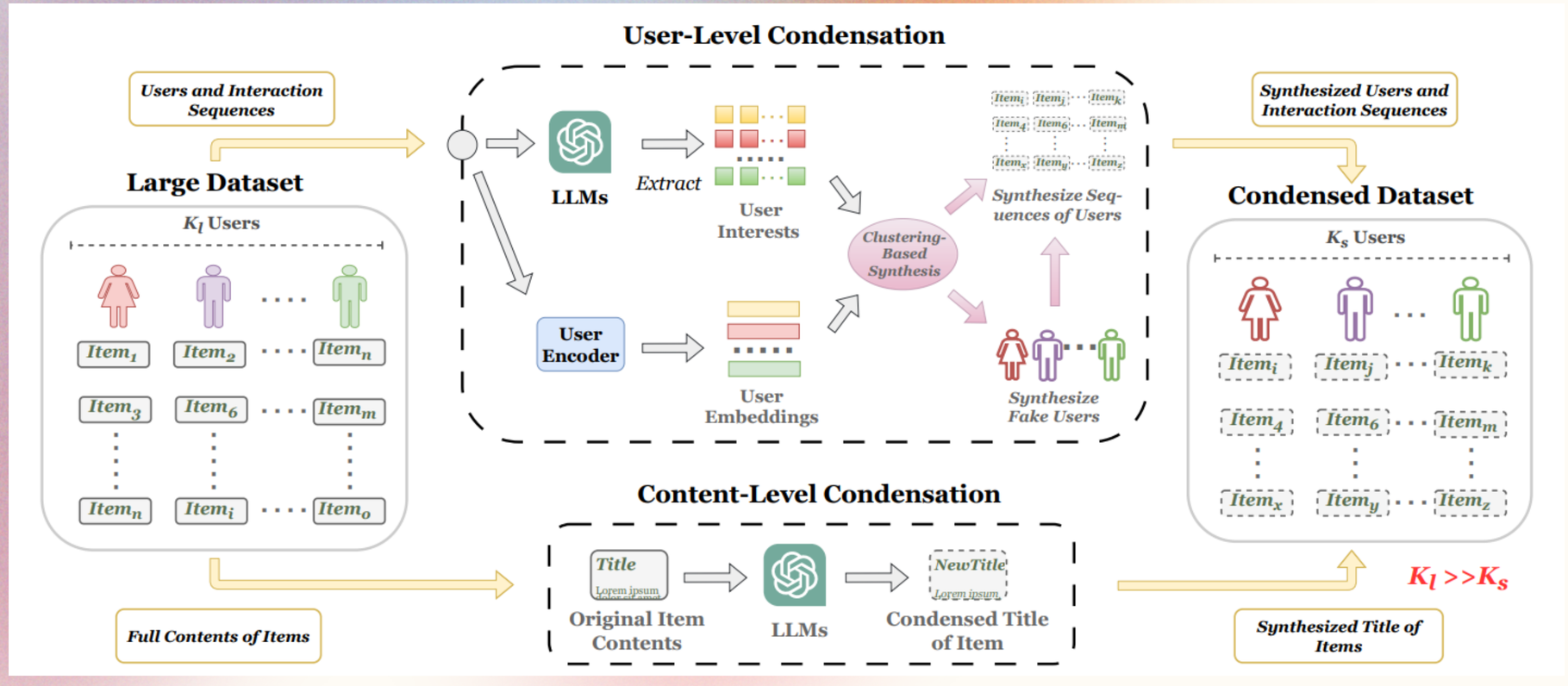
Why LLM's?

Firstly they mainly focused on textual data as there are no existing methods that deal with it and given the fact that text data is extremely important in Recommendation systems. Another crucial aspect is maintaining the relationship between user and item.

Architecture

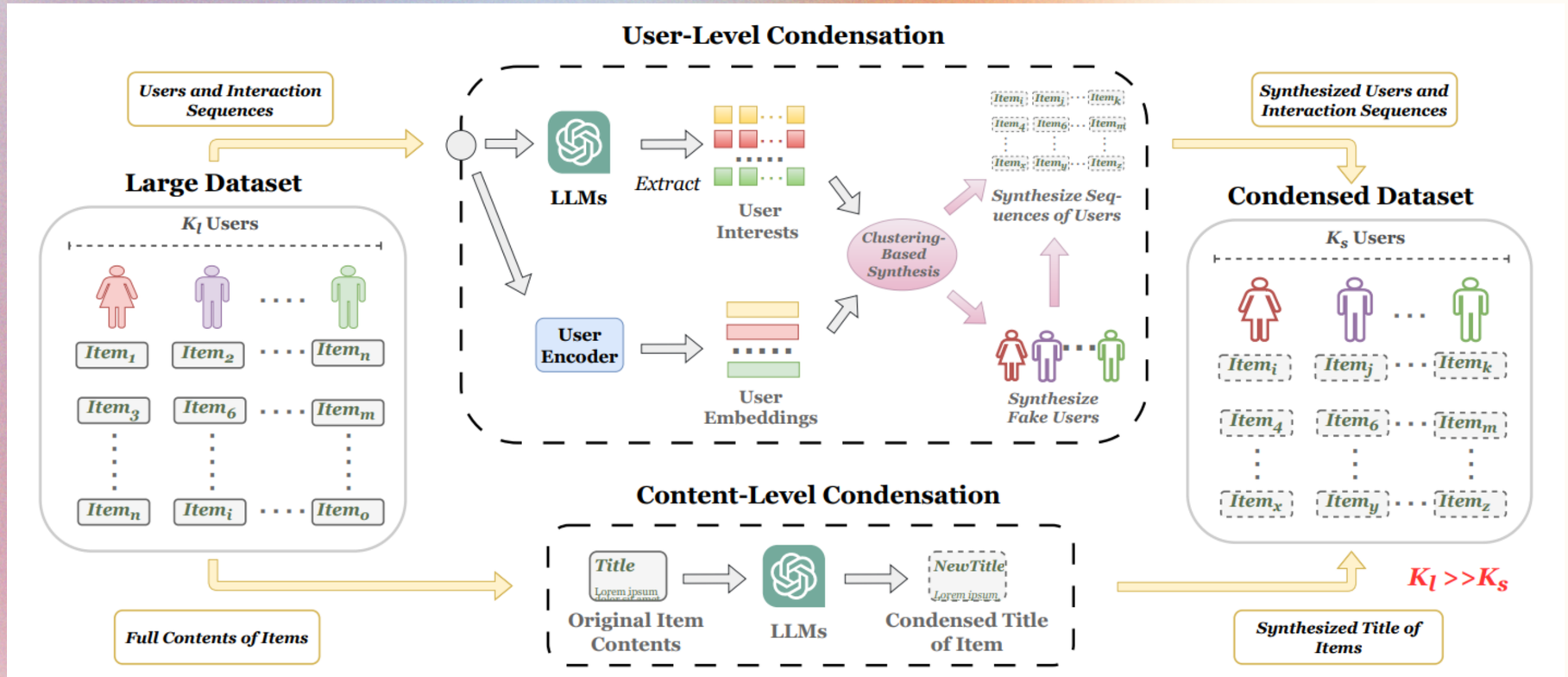
Two Levels:

- User Level
- Content Level



Content Level Condensation

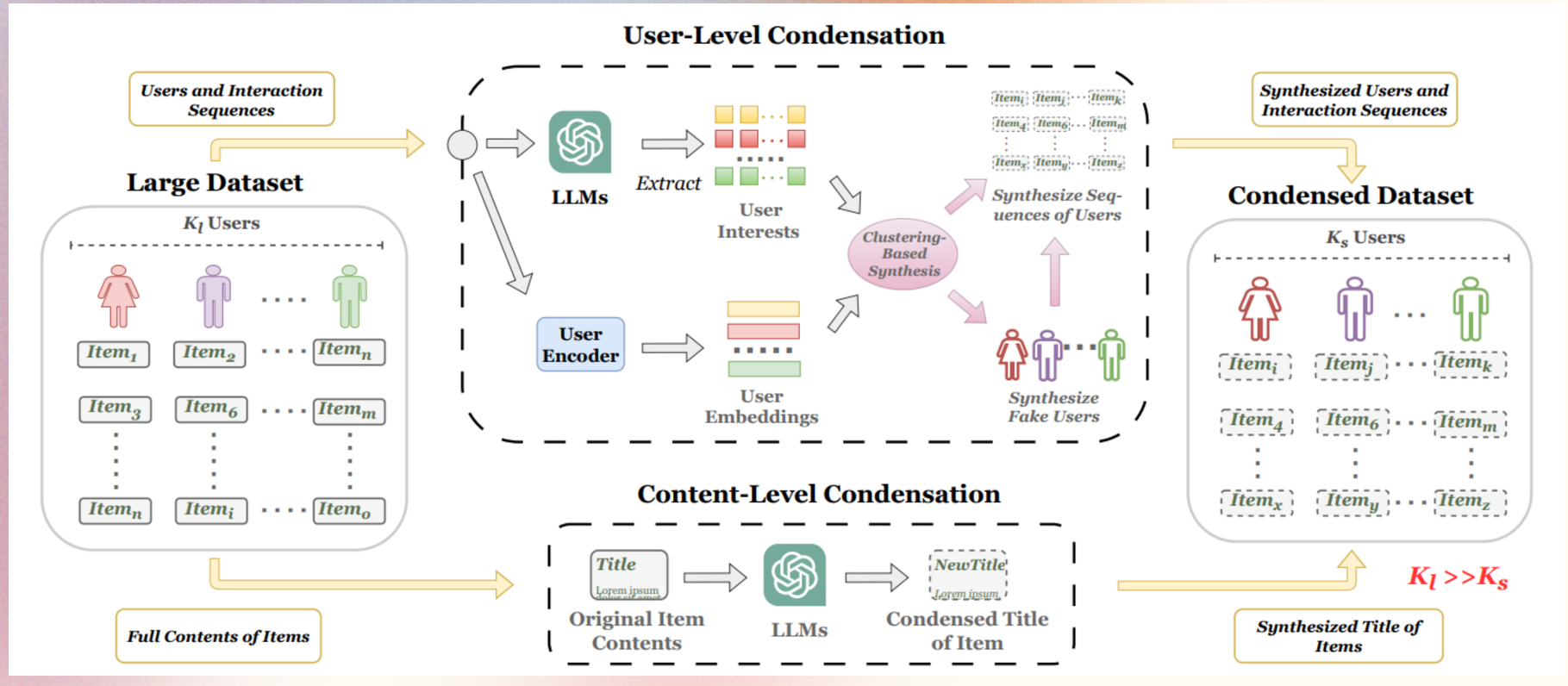
$$n_s = LLM(n),$$



User Level Condensation

$$z_u = f_\theta(h^{(u)}, \mathcal{N}),$$

$$d_u^{emb} = Dis(z_u, c_i),$$

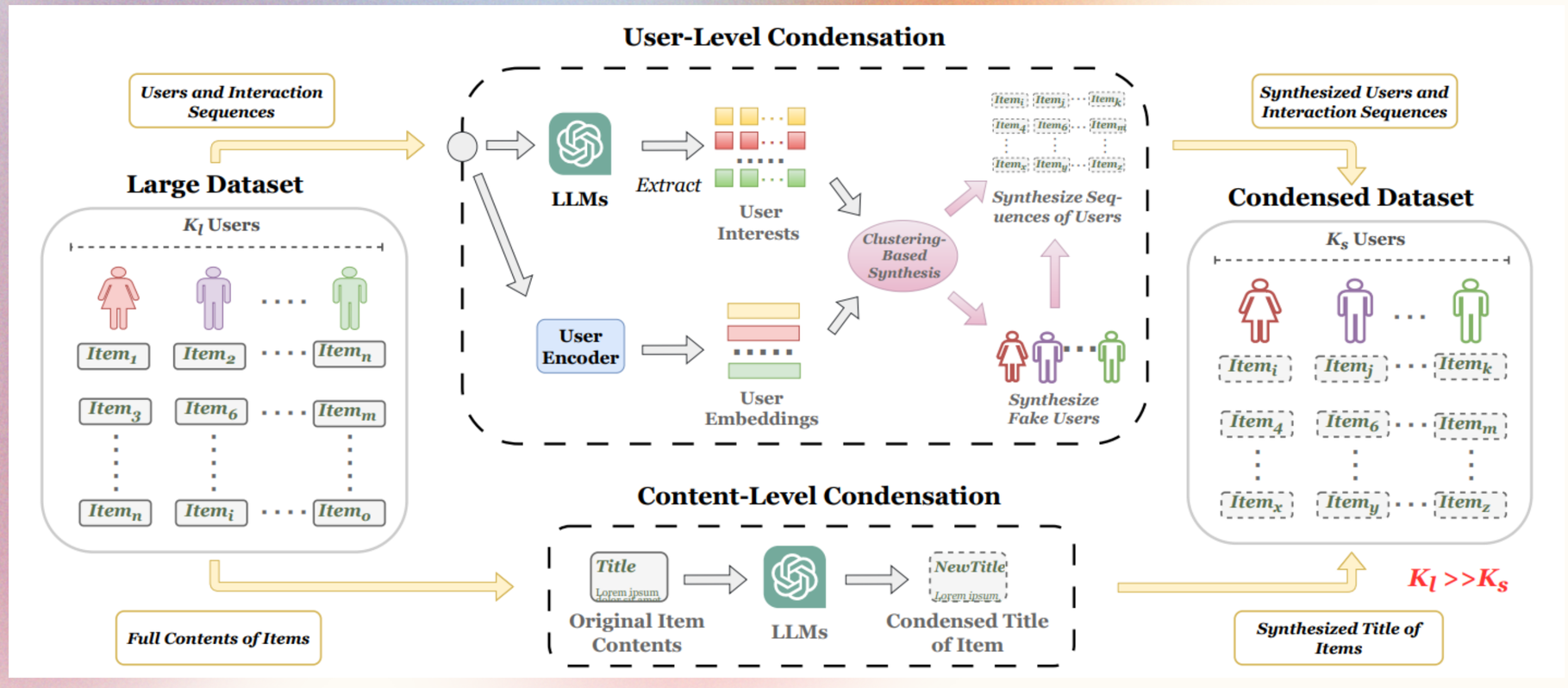


User Level Condensation

$$\mathcal{I}^{(u)} = LLM(h^{(u)})$$

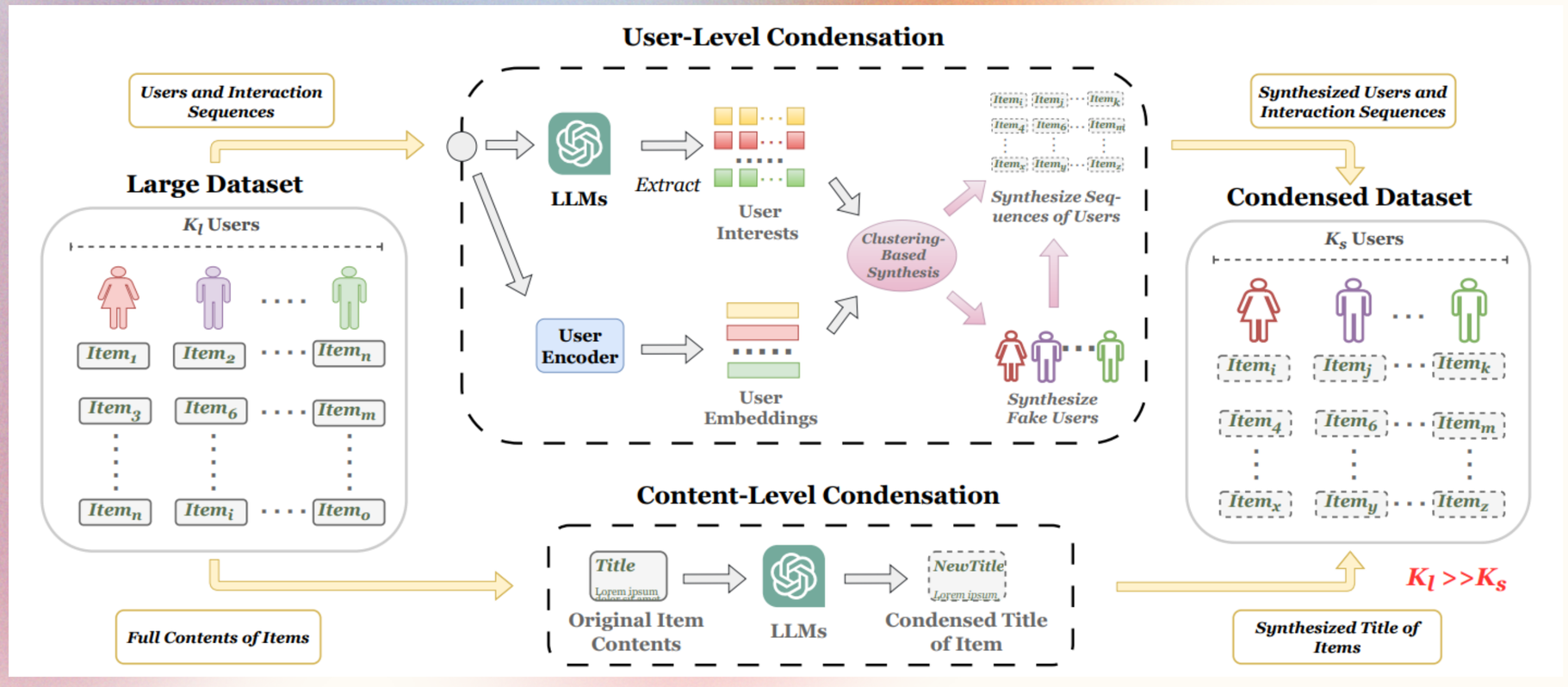
$$e_u = f_{pool}\left(f_{PLM}(\mathcal{I}^{(u)})\right)$$

$$d_u^{int} = Dis(e_u, c_i').$$



User Level Condensation

$$d_u = d_u^{emb} + \alpha \cdot d_u^{int}$$



LLM Interactions Examples

Enhance item titles based on given contents in the following format:

[title] {title}, [abstract] {abstract}, [category] {category}

You should rephrase the title to be clear, complete, objective, and neutral. Only provide the new title in the following format:

[newtitle] {newtitle}

[title] {Health Weightloss Watch},

[abstract] {Man Shares Time-Lapse Video of Six-Month Weight-Loss Journey We're big fans of weight-loss stories, but we usually only get to see the before and after photos. Very rarely do we get to see someone's physique transform right before our very eyes.},

[category] {Health}

[newtitle] {A Six-Month Weight-Loss Journey Captured in Time-Lapse Video},

Describe user's interests his/her browsed lists, which will be input in the following format:

(1) {title}, (2) {title}, ..., (n) {title}

You an only response user's interests in following format:

[interests] - interest1 - interest2 - interest3 .....

(1) Chrisleys sue Georgia tax official with link to estranged daughter

(2) Former President Jimmy Carter suffers pelvic fracture after falling in Georgia home

(3) This girl nails the wave pool then fails the exit spectacularly

.....

[interests] - celebrity drama - sports .....

Metrics Used in Paper

NDCG@K: Measures the quality of the recommended items by considering both the position of the relevant items in the list and their individual relevance scores.

Recall@K: Measures the ability of the system to retrieve a set of relevant items out of all possible relevant items in the dataset.

Quality: Measure the performance of the condensed dataset considering the performance on original dataset is 100%

Ratio: Dataset compression ratio with respect to original dataset

Results

Datasets		MIND				Goodreads				MovieLens			
Methods		**OR.**	**RD.**	**MJ.**	**Ours**	**OR.**	**RD.**	**MJ.**	**Ours**	**OR.**	**RD.**	**MJ.**	**Ours**
Item	**avg. tok.**	45.42	16.73	16.73	16.73	35.37	16.01	16.73	16.01	34.80	15.26	16.73	15.26
	Size (KB)	10,384	4,095	4,095	4,095	2,189	993	993	993	232	121	121	121
	Ratio	100%	39%	39%	39%	100%	45%	45%	45%	100%	52%	52%	52%
User	**# users**	94,057	9,405	9,405	9,405	23,089	4,617	2,350	4,617	943	47	47	47
	Size (KB)	106,614	2,124	5,733	2,139	8,704	1,791	1,977	1,957	532	68	232	81
	Ratio	100%	2%	5%	2%	100%	21%	23%	22%	100%	13%	44%	15%
Overall	**Size (KB)**	116,998	6,219	9,828	6,234	10,893	2,784	2,970	2,950	764	189	353	202
	Ratio	100%	5%	8%	5%	100%	26%	27%	27%	100%	25%	46%	26%

Datasets		MIND$^!$, r=5%				Goodreads, r=27%				MovieLens, r=26%			
Rec Model	**Metrics**	**Random**	**Majority**	**TF-DCon**	***Original***	**Random**	**Majority**	**TF-DCon**	***Original***	**Random**	**Majority**	**TF-DCon**	***Original***
NAML	**N@1**	0.2871	0.2854	0.3071	*0.3176*	0.5197	0.5057	0.5411	*0.6462*	0.8241	0.8397	0.8484	*0.828*
	N@5	0.3470	0.3466	0.3691	*0.3783*	0.7943	0.7884	0.8033	*0.8475*	0.8251	0.8367	0.8494	*0.831*
	R@1	0.4016	0.4002	0.4377	*0.4534*	0.4520	0.4326	0.4704	*0.5635*	0.1785	0.1886	0.1873	*0.1752*
	R@5	0.5670	0.5697	0.6150	*0.6270*	0.9983	0.9986	0.9984	*0.9989*	0.7274	0.7323	0.7475	*0.7399*
	Quality	90.28%	90.15%	**97.22%**	*100.00%*	88.57%	87.01%	**90.49%**	*100.00%*	99.75%	102.18%	**103.15%**	*100.00%*
NRMS	**N@1**	0.2625	0.2631	0.2997	*0.3009*	0.5399	0.5094	0.5453	*0.6439*	0.8105	0.8294	0.8149	*0.8178*
	N@5	0.3225	0.3225	0.3597	*0.3608*	0.8017	0.7901	0.8054	*0.8476*	0.8227	0.8334	0.8371	*0.8253*
	R@1	0.3750	0.3793	0.4279	*0.4325*	0.4704	0.4344	0.4751	*0.5629*	0.1729	0.1821	0.1798	*0.1757*
	R@5	0.5414	0.5445	0.6000	*0.6042*	0.9982	0.9990	0.9985	*0.9993*	0.7325	0.7339	0.7446	*0.7321*
	Quality	88.23%	88.66%	**99.38%**	*100.00%*	90.47%	87.37%	**91.01%**	*100.00%*	99.31%	**101.57%**	101.28%	*100.00%*
Fastformer	**N@1**	0.2815	0.2736	0.3022	*0.3057*	0.5420	0.5165	0.5548	*0.6556*	0.7886	0.8105	0.8251	*0.7915*
	N@5	0.3425	0.3350	0.3637	*0.3645*	0.8028	0.7934	0.8092	*0.8529*	0.8254	0.8318	0.8429	*0.8145*
	R@1	0.3944	0.3804	0.4334	*0.4365*	0.4725	0.4414	0.4851	*0.5745*	0.1738	0.1799	0.1836	*0.166*
	R@5	0.5631	0.5515	0.6096	*0.6144*	0.9983	0.9991	0.9986	*0.9990*	0.739	0.7373	0.7465	*0.7279*
	Quality	92.01%	89.58%	**99.29%**	*100.00%*	89.74%	87.16%	**90.97%**	*100.00%*	101.80%	103.55%	**105.22%**	*100.00%*

My two cents on paper

1. The paper really presented the novel approach for Dataset condensation which is extremely useful in Recommendation Systems these days. The way they leveraged LLM in solving this complex task is really awesome. Towards the end they used two fundamental techniques: 1)KMeans - clustering and 2) LLM for text condensation which is what most of us know but the greatness comes in the way they leveraged it.
2. The results or metrics they presented are really promising and it is really gonna revolutionize the Recommendation Systems. With just like subset of dataset they achieved nearly same or sometimes even better results

Q&A Session

Thank you for listening!

References

- Leveraging Large Language Models (LLMs) to Empower Training-Free Dataset Condensation for Content-Based Recommendation