Mon, Sept 18 2023

Politecnico di Bari

ADVISOR

Tommaso Di Noia
Vito Walter Anelli

PH.D. STUDENT

Dario Di Palma

Retrieval-augmented Recommender System: Enhancing Recommender Systems with Large Language Models

17th ACM Conference on Recommender Systems, Singapore, 18th-22nd September 2023

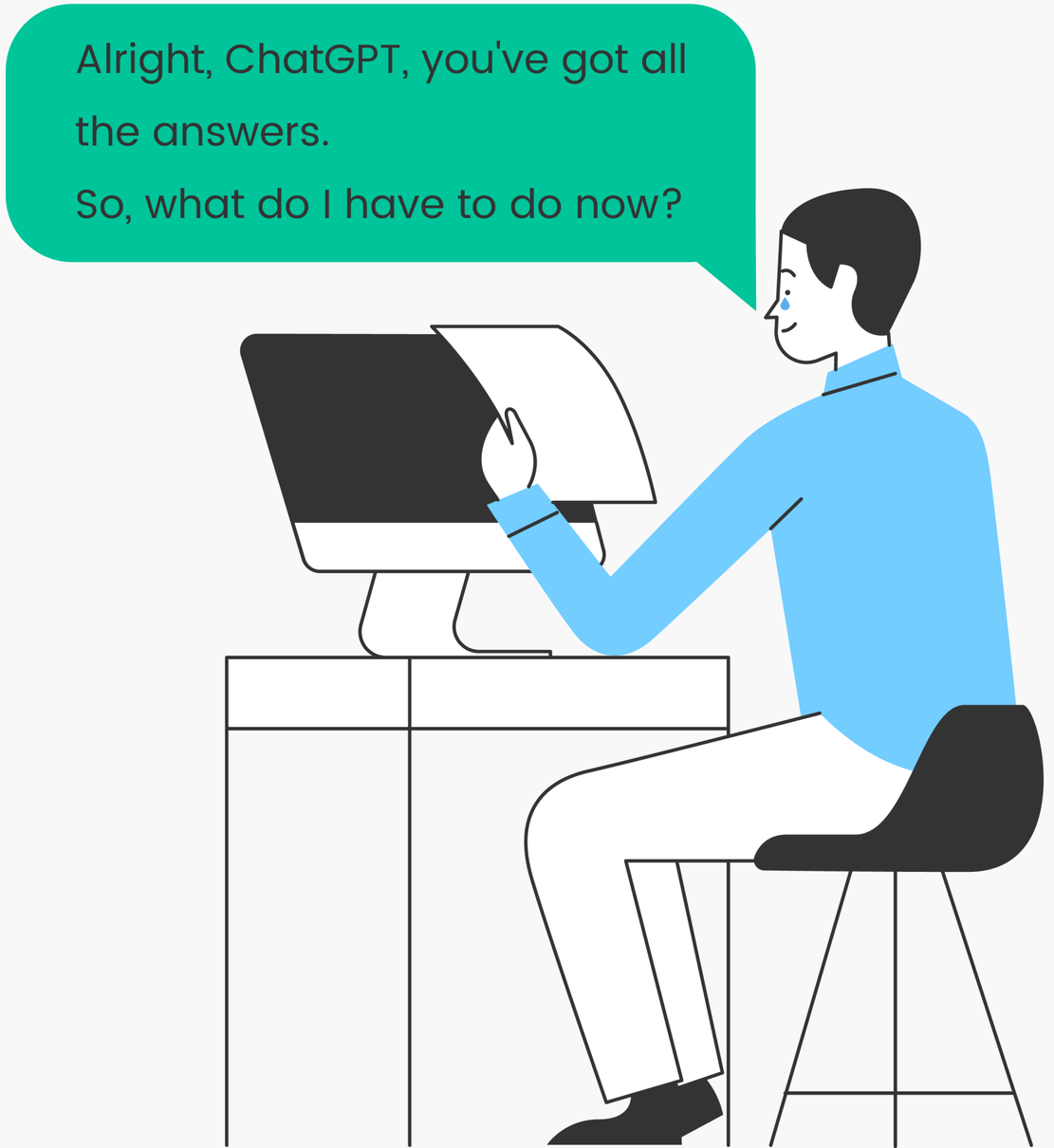
I Research Background & Motivation

- 30 November 2022 -> Unveiling ChatGPT to the world.

First year Phd. Student

I Research Background & Motivation

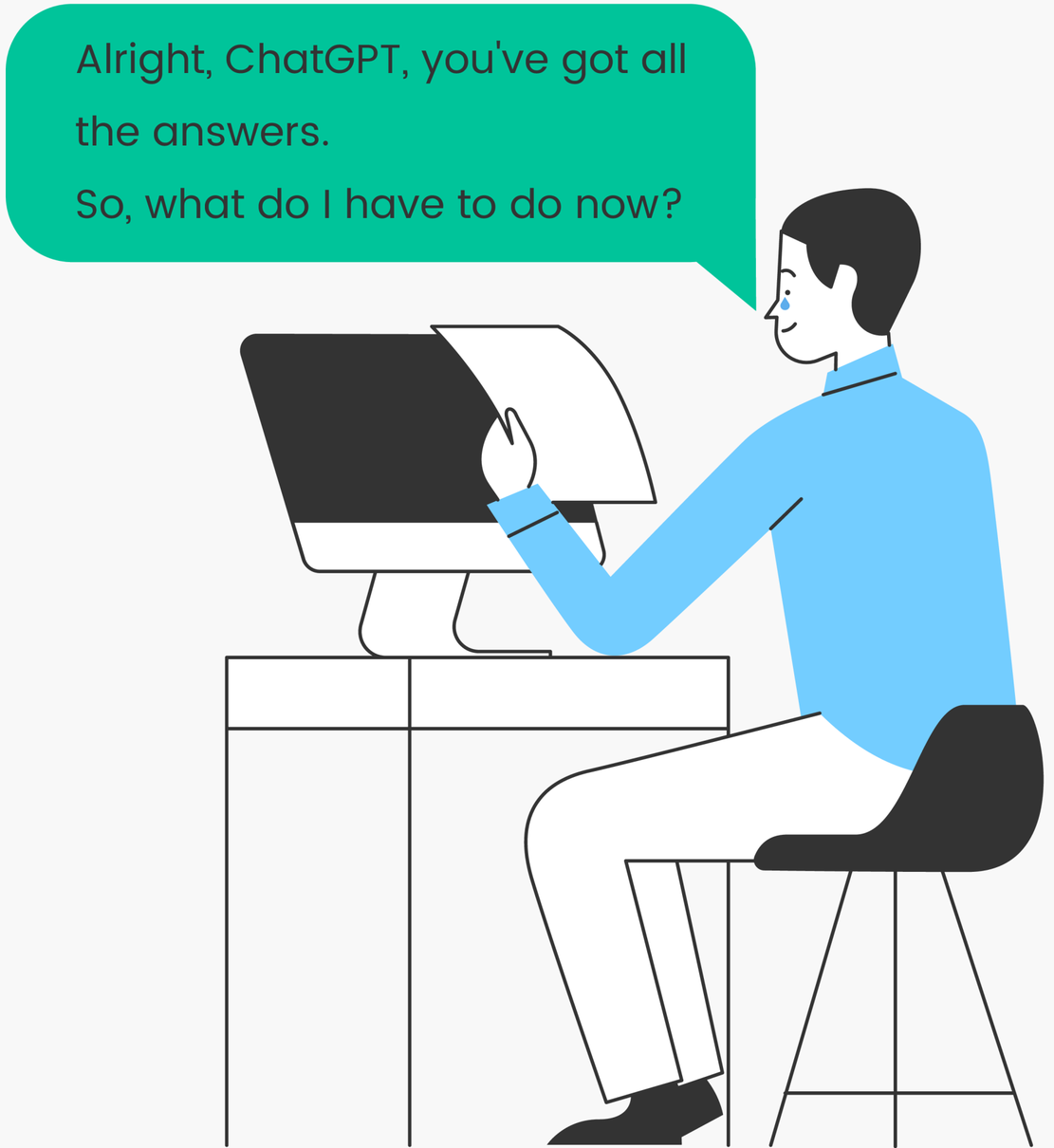
- 30 November 2022 -> Unveiling ChatGPT to the world.



I Research Background & Motivation

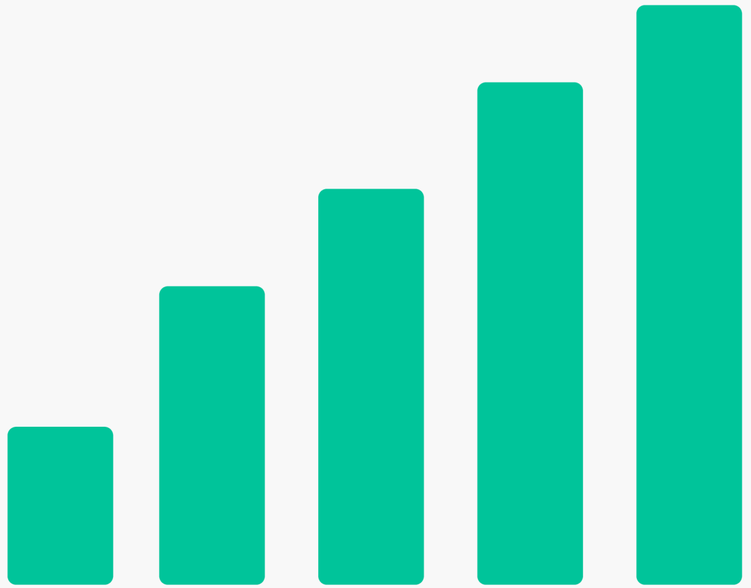
- 30 November 2022 -> Unveiling ChatGPT to the world.

- However, from an in-depth study, it's clear that ChatGPT is only one thing:
  - an exceptional assistant capable of handling a wide range of tasks.



The Rise of Open Large Language Models

- Llama 1 & 2 - Meta
- MPT - MosaicML
- Falcon - TII
- Vicuna - LMSYS Org



“A Large Language Model is a type of artificial intelligence (AI) system that has been trained on vast amounts of text data to understand and generate human language.”
- ChatGPT

How proficient are Language Models, such as ChatGPT, on the Recommendation task? Zero-Shot Scenario

"Given a user, as a recommender system, provide recommendations. The user [X] likes the following [items]: [item_1], [item_2], etc. Give me back 50 recommendations."

MovieLens	
Model	nDCG@10
UserKNN	**0.32358**
ItemKNN	0.31702
ChatGPT-3.5	0.16927

Facebook Books	
Model	nDCG@10
ChatGPT-3.5	**0.05742**
AttributeItemKNN	0.05034
VSM	0.04592

How proficient are Language Models, such as ChatGPT, on the Recommendation task? Zero-Shot Cold-Start Scenario

"Given a user, as a recommender system, provide recommendations. The user [X] likes the following [items]: [item_1], [item_2], etc. Give me back 50 recommendations."

MovieLens	
Model	nDCG@10
PaLM-2	**0.14032**
ChatGPT-3.5	0.08719
RP3β	0.03052

Facebook Books	
Model	nDCG@10
ChatGPT-3.5	**0.04871**
PaLM-2	0.03975
EASER	0.00918

How proficient are Language Models, such as ChatGPT, on the Recommendation task?

Nevertheless, to **fully comprehend** the effectiveness of **LLMs in** the **Recommendation task**, it is necessary to **conduct** further investigations into their performance using **prompt engineering** techniques such as **Chain of Thought or Tree of Thought**.

What are the limitations of using LLMs for recommendations?

Hallucinations

if not
Limited

Knowledge Cutoff

What are the limitations of using LLMs for recommendations?

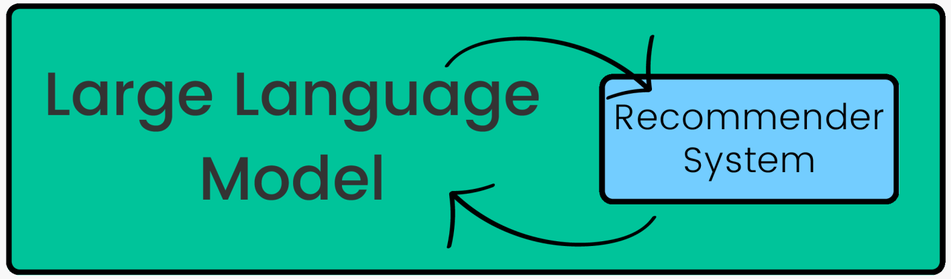
Research Directions in LLMs:

- Pre-training
- Prompt Engineering
- Fine-tuning

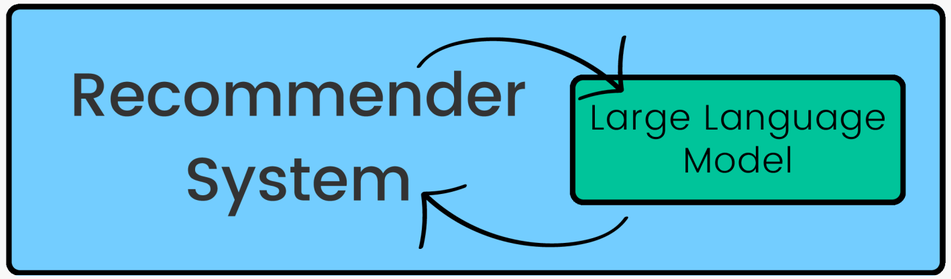
Research Directions in RSs:

- Preference Acquisition
- Interaction
- New Recommendation Tasks

How can the combination of retrieval-based (RS) and generation-based (LLM) methods improve the quality of the recommendations?

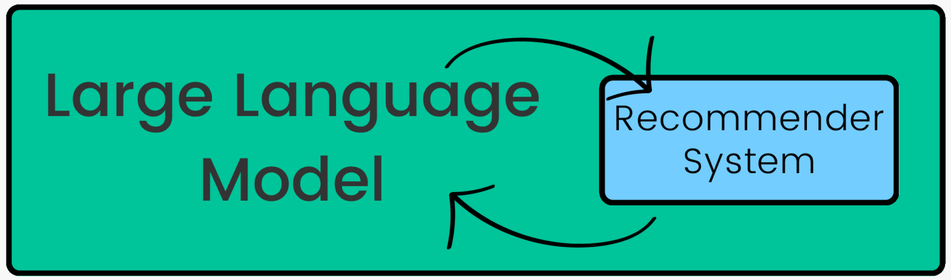


Top-down approach: LLMs lead the dialogue, while RS step in to offer valuable recommendations.
› **Conversational Recommender Systems**.



Bottom-up approach: RS provide recommendations while the LLM is a plug-in designed to enhance those recommendations.
› **Retrieval-augmented Recommender System**.

How can the combination of retrieval-based (RS) and generation-based (LLM) methods improve the quality of the recommendations?

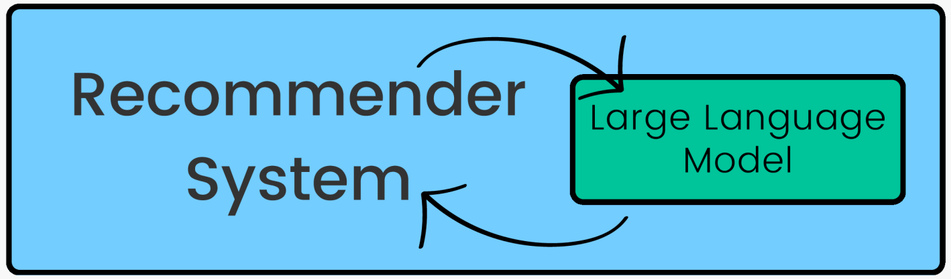


Top-down approach: LLMs lead the dialogue, while RS step in to offer valuable recommendations.
> **Conversational Recommender Systems**.

Example of Top-down approach:

- ChatRec[Gao et al. 2023] > converting user-profiles and history into prompts.
- GenRec[Wang et al. 2023] > AI content generator from user instructions
- BookGPT[Zhiyuli et al. 2023] > a ChatGPT-like book recommendation system

How can the combination of retrieval-based (RS) and generation-based (LLM) methods improve the quality of the recommendations?



Bottom-up approach: RS provide recommendations while the LLM is a plug-in designed to enhance those recommendations.
> **Retrieval-augmented Recommender System**.

Large Language Models for:

- Extracting Item Features.
- Data Augmentation.
- Providing Explanations.
- Developing Methods for User and Item Representation.
- Modelling User Behavior through Explicit Preference Acquisition.

II Discussions & Feedback

- What is the community's perspective on the integration of LLMs into RSs?
- How can we ensure the reproducibility of our work in the context of testing various combinations of LLMs and RSs?
- Can you identify any specific challenges or limitations encountered when utilizing large language models for these purposes, excluding computational power considerations?
- Were there instances where the augmented data appeared unnatural or introduced biases during the integration process?
- In your expert opinion, what improvements or enhancements do you recommend for optimizing the utilization of large language models in these applications?
- Did you encounter any resistance or face particular challenges in gaining user trust and acceptance of model-generated recommendations or explanations?

Mon, Sept 18 2023

Politecnico di Bari

MOST
CENTRO NAZIONALE PER LA MOBILITÀ SOSTENIBILE

SISINF LAB

ADVISOR

Tommaso Di Noia
Vito Walter Anelli

PH.D. STUDENT

Dario Di Palma

Thank you for listening!

d.dipalma2@phd.poliba.it

RecSys
acm.org

Save Contact!