(A short intro to) LLMs-based Recommender Systems

Marie Al-Ghossein

Crossing Minds

June 2, 2025

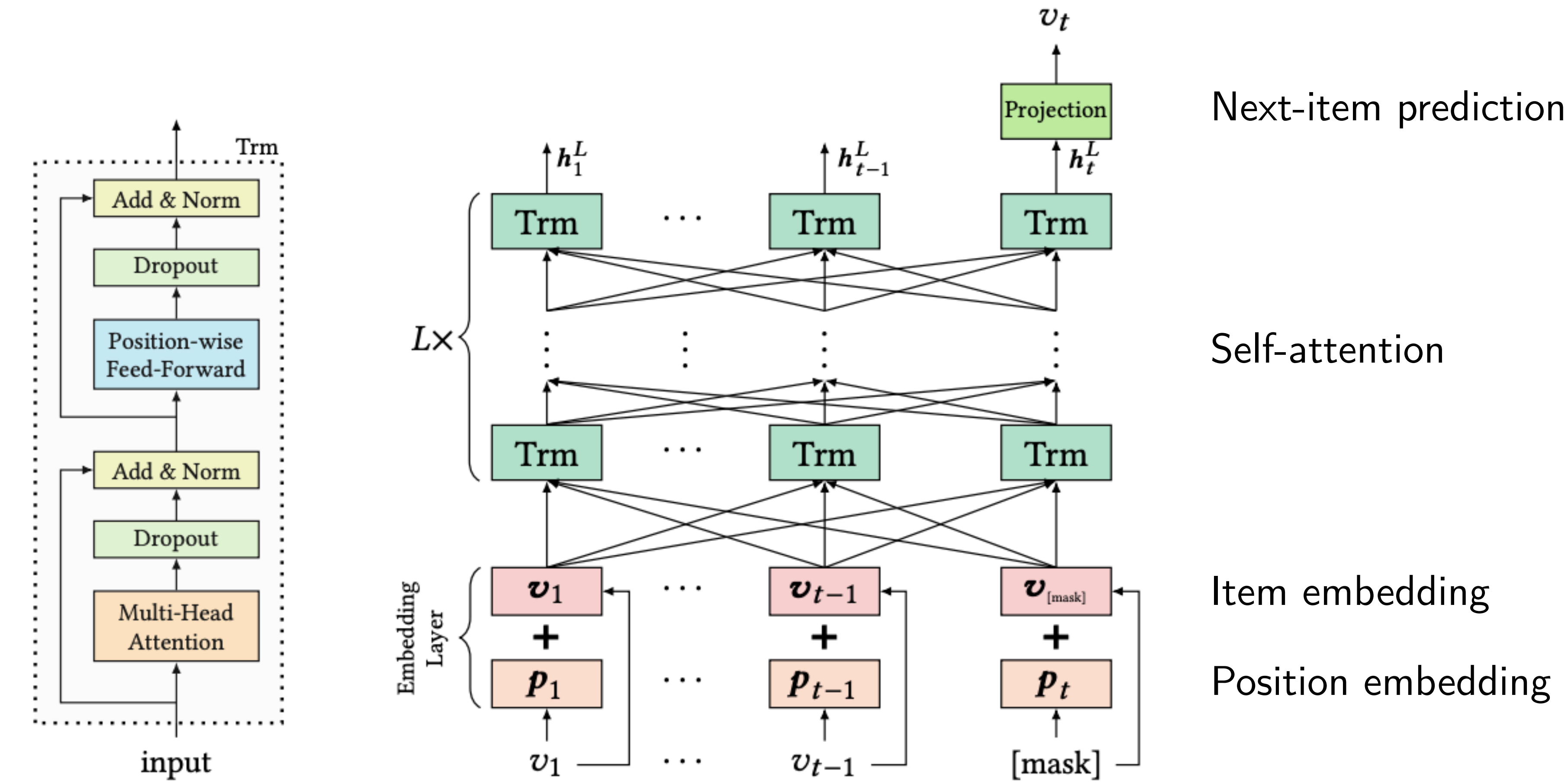
The emergence of Large Language Models (LLMs)

- LLMs emerging as a powerful tool for NLP
- Remarkable abilities in several areas, e.g.,
  - Text completion, summarization, Q&A, translation, among others
- Billions of parameters, trained on a chunk of the internet
- Pretrained models, supports fine-tuning and in-context learning

Claude 3 | Olympus | PaLM-2 | Gemini | Llama-2 | Large | GPT 4

Generative Models for RecSys

BERT4Rec for sequential recommendation



(a) Transformer Layer.

(b) BERT4Rec model architecture.

NLP:

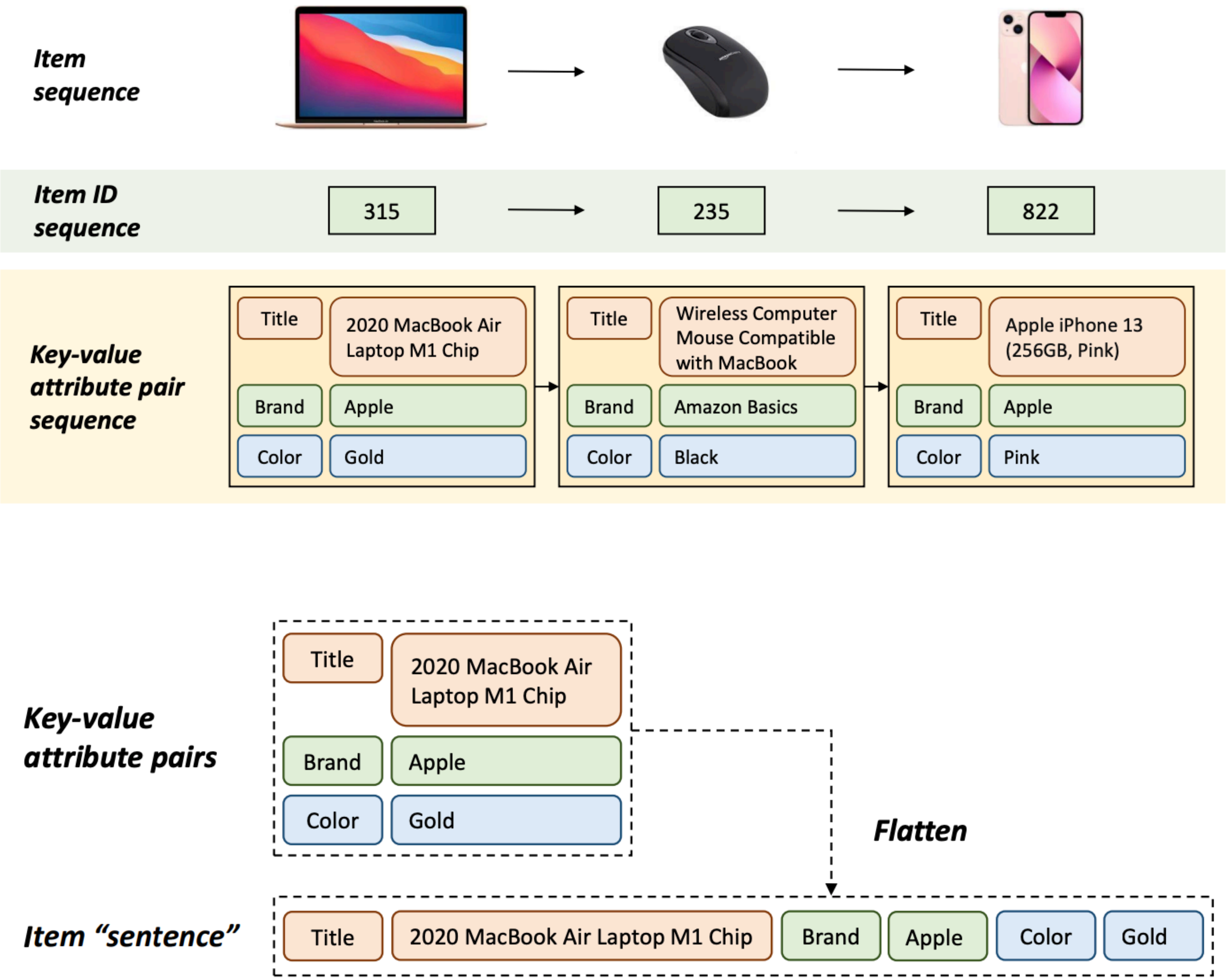
- Token sequence
- Inter-token correlations

RecSys:

- ID sequence
- Inter-item correlations

Sun et al., "BERT4Rec: Sequential Recommendation with Bidirectional Encoder Representations from Transformer", CIKM 2019

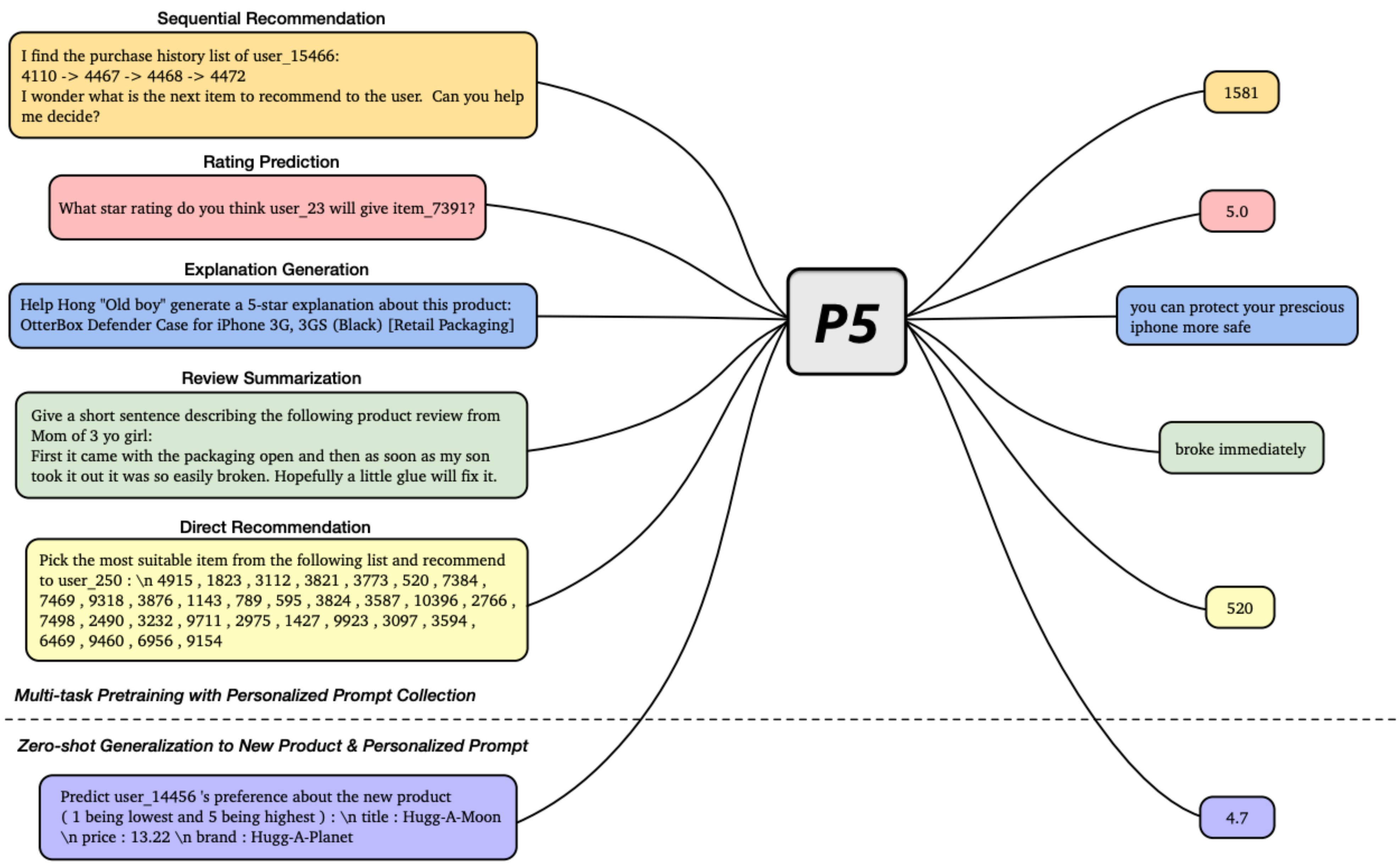
RecFormer - Text is all you need



Li et al., "Text Is All You Need: Learning Language Representations for Sequential Recommendation", KDD 2023

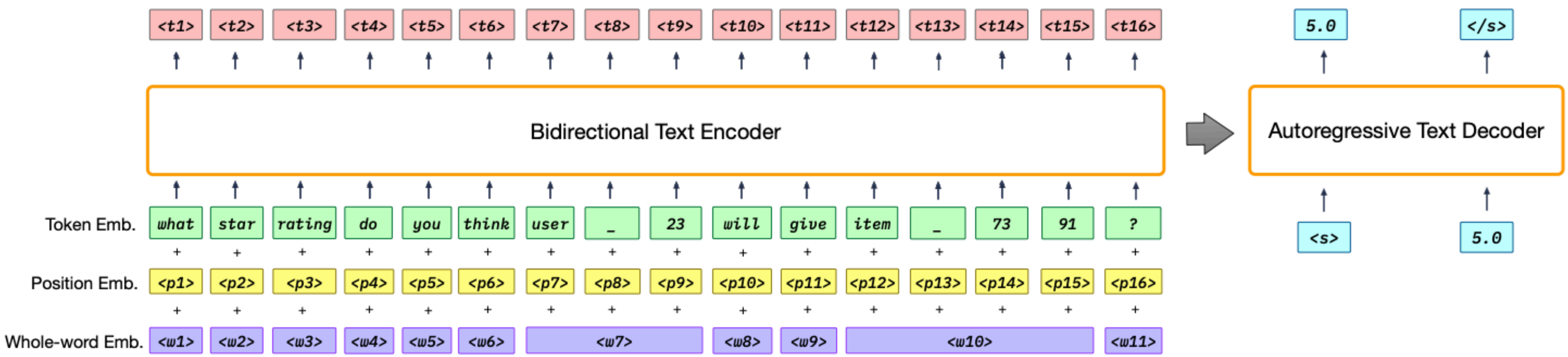
In-Context Learning/LLMs for RecSys

Recommendation as language processing



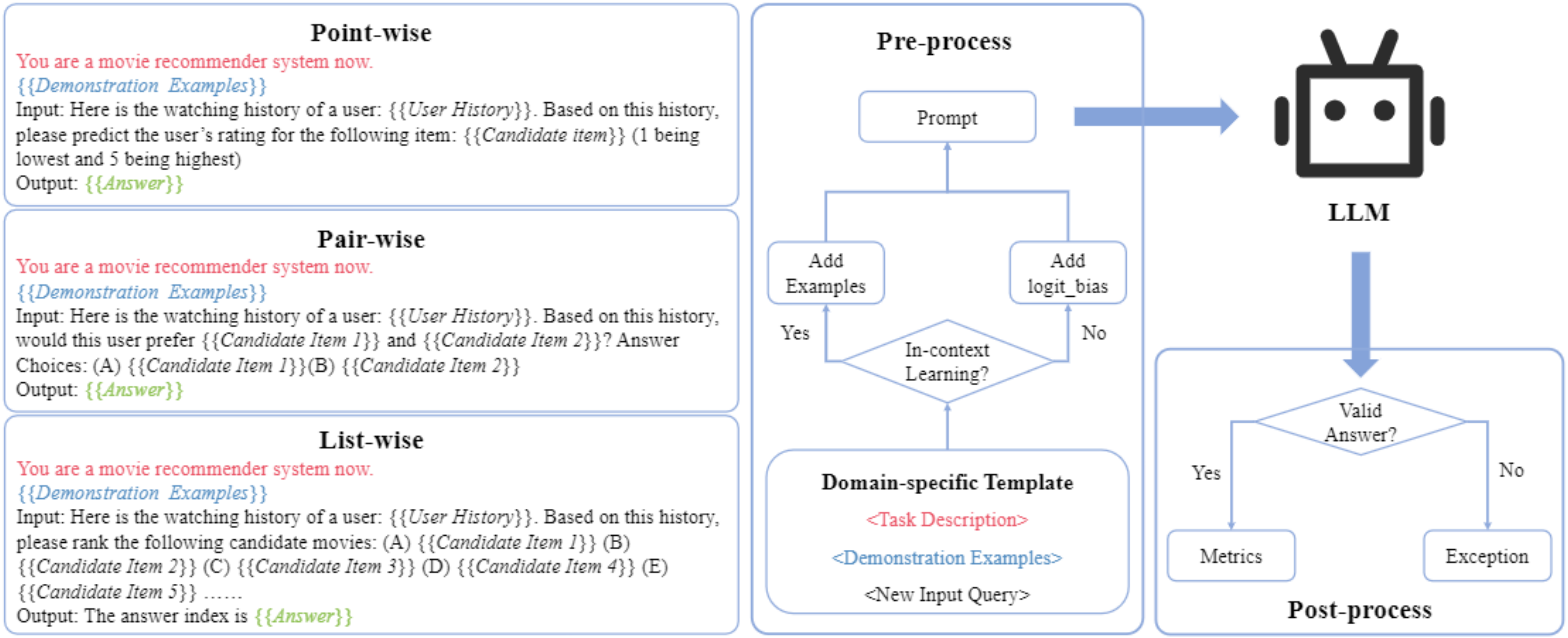
Geng et al., "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)", RecSys 2022

Recommendation as language processing



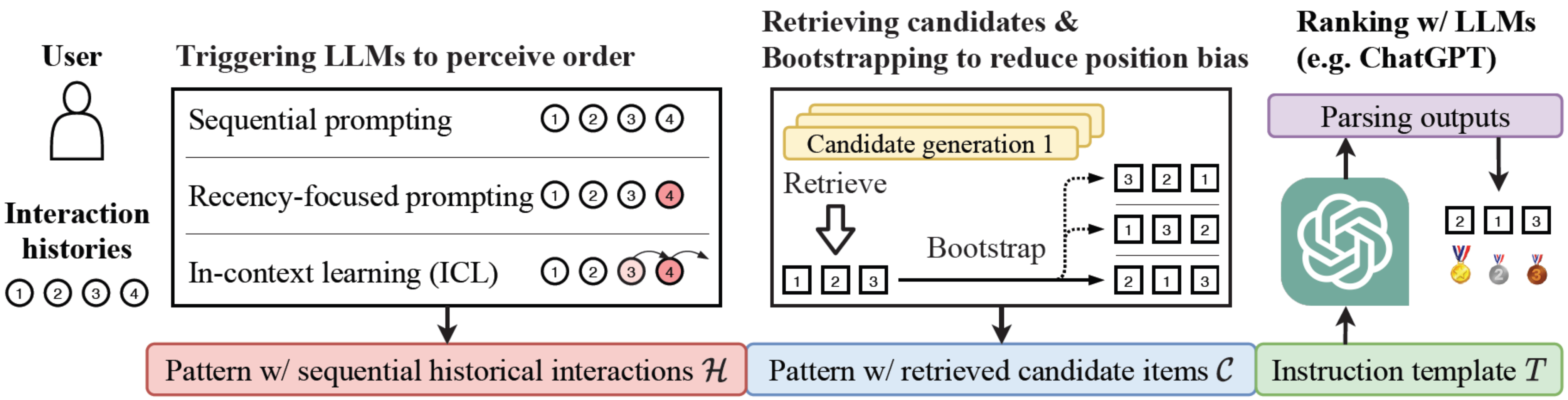
Geng et al., "Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm (P5)", RecSys 2022

LLMs as a reranker



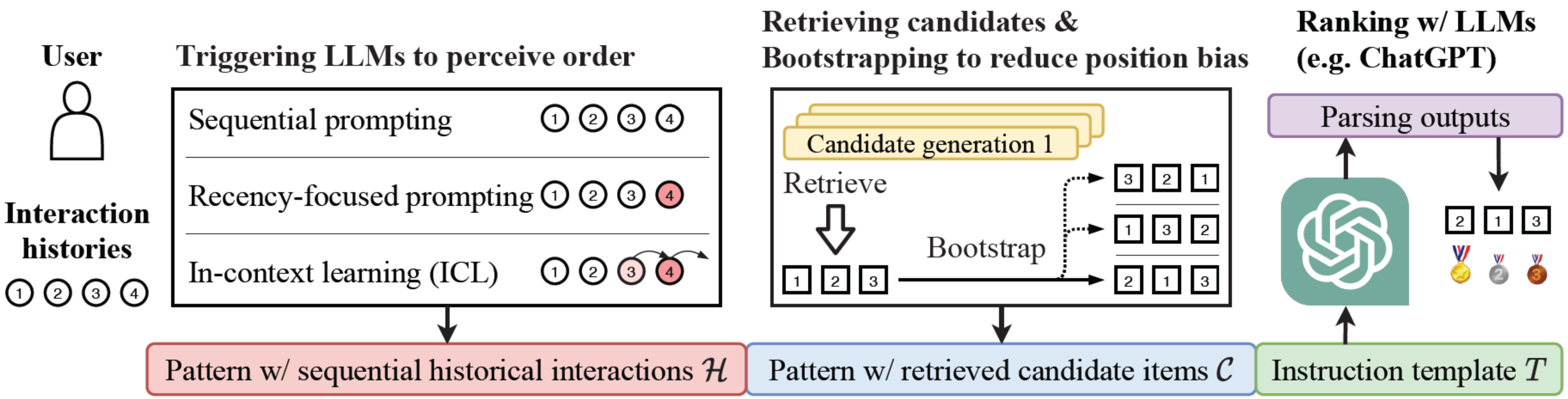
Dai et al., "Uncovering ChatGPT's Capabilities in Recommender Systems", RecSys 2023

Addressing position bias



Hou et al., "Large Language Models are Zero-Shot Rankers for Recommender Systems", ECIR 2024

Addressing position bias



Triggering LLMs to perceive order

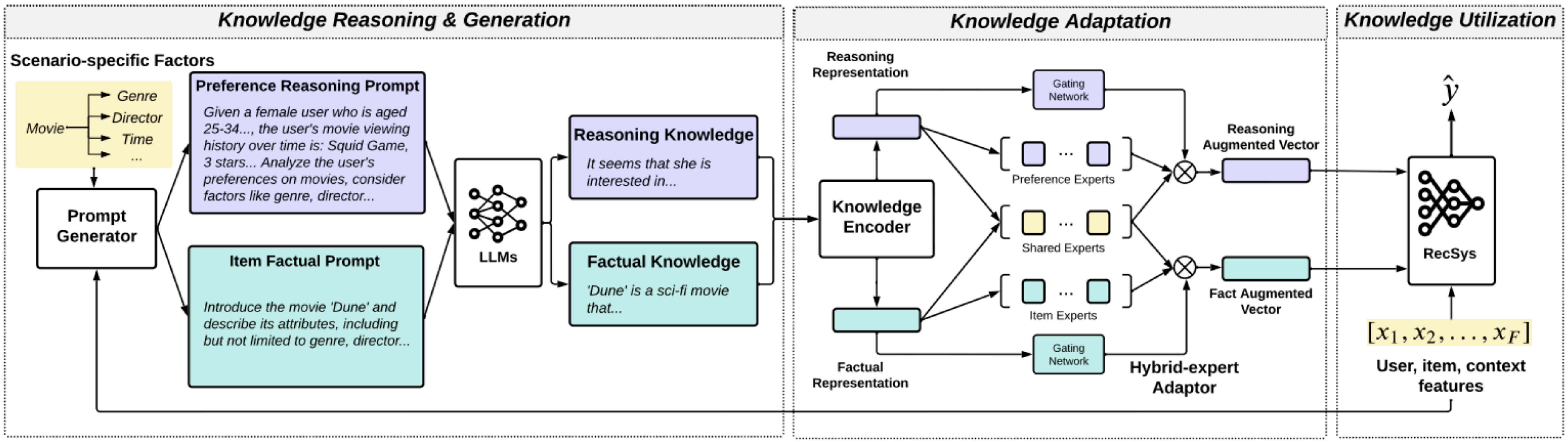
- Sequential prompting: Historical interactions in their natural order
- Recency-focused prompting: *"Note that my most recently watched movie is [...]"*
- In-context learning: *"If I've watched the following movies in the past in order: [...] then you should recommend [...] to me and now that I've watched [...], then:"*

Bootstrapping candidates

- Randomly shuffle candidates, use each sample to query the LLM, combine outputs

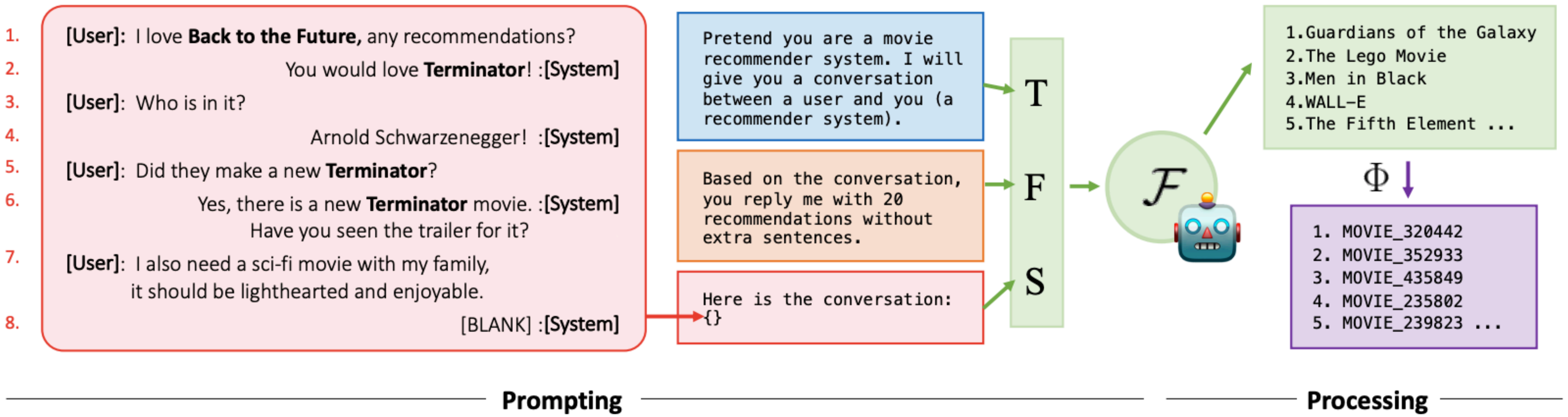
Hou et al., "Large Language Models are Zero-Shot Rankers for Recommender Systems", ECIR 2024

Knowledge augmentation from LLMs



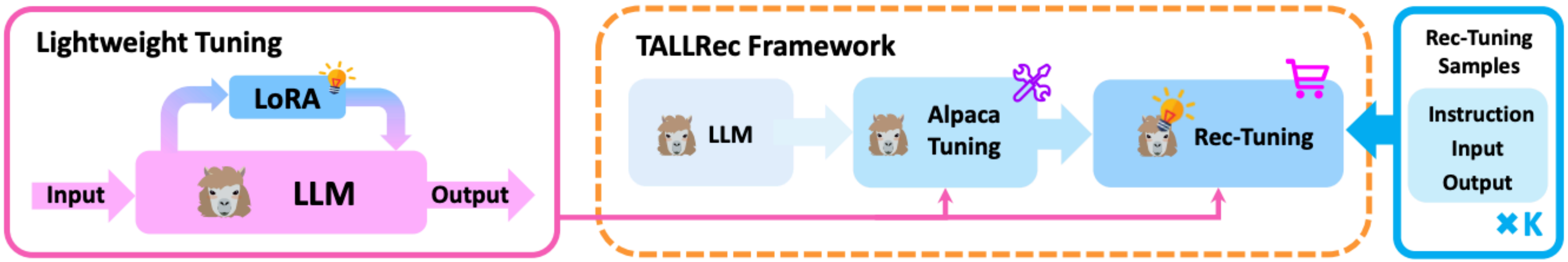
Xi et al., "Towards Open-World Recommendation with Knowledge Augmentation from Large Language Models", arXiv 2023

LLMs as zero-shot conversational RS



He et al., "Large Language Models as Zero-Shot Conversational Recommenders", CIKM 2023

Tuning LLMs for recommendation



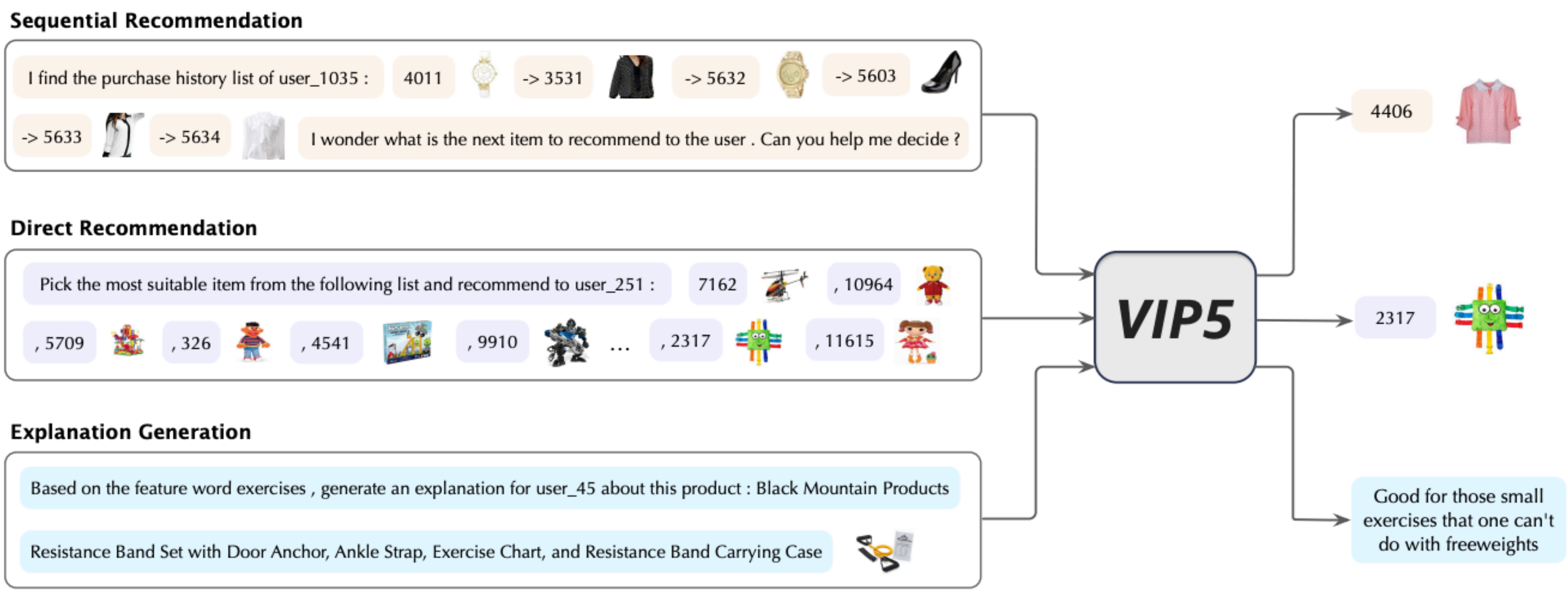
Transforms the recommendation data as instructions used to tune the LLM via an instruction tuning process, e.g.,

Instruction Input	
Task Instruction:	Given the user's historical interactions, please determine whether the user will enjoy the target new movie by answering "Yes" or "No".
Task Input:	User's liked items: GodFather. User's disliked items: Star Wars. Target new movie: Iron Man
Instruction Output	
Task Output:	No.

Bao et al., "TALLRec: An Effective and Efficient Tuning Framework to Align Large Language Model with Recommendation", RecSys 2023

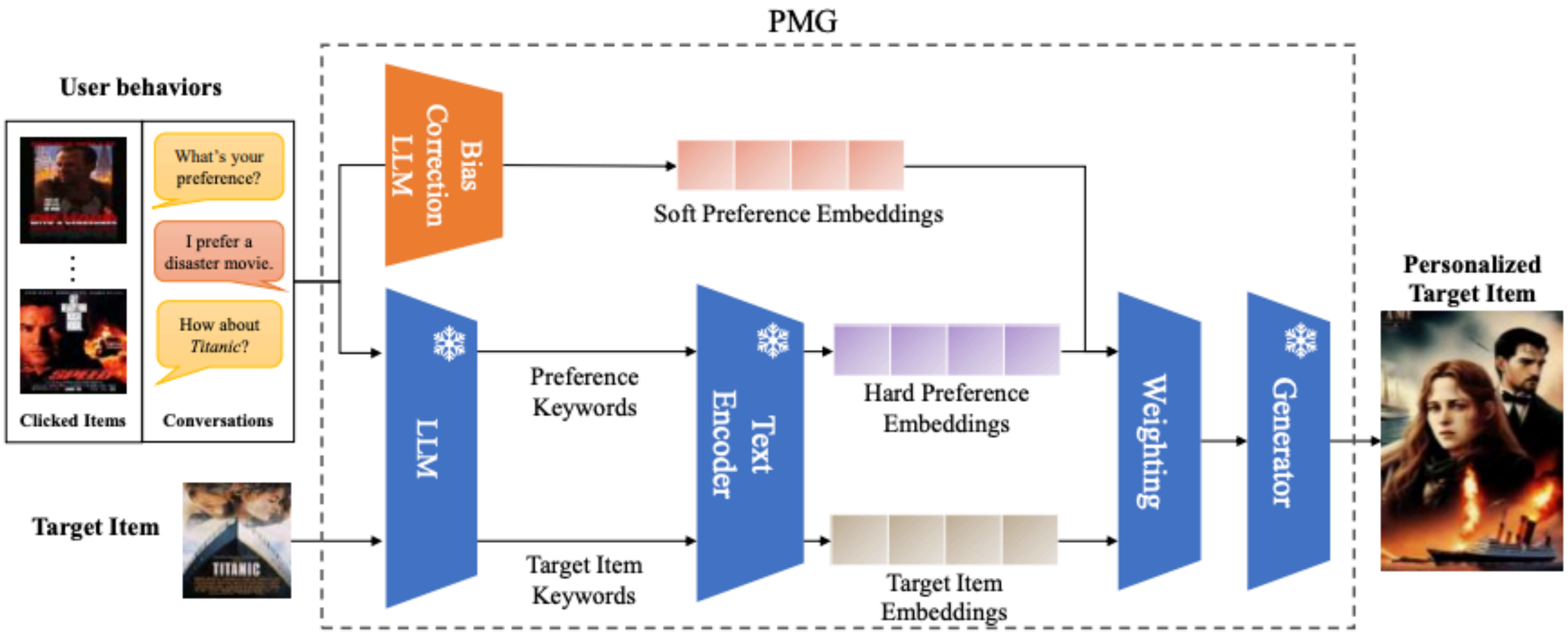
Multimodal RecSys

Multimodal foundation models for recommendation



Geng et al., "VIP5: Towards Multimodal Foundation Models for Recommendation", EMNLP 2023

Personalized multimodal generation with LLMs



(a) $w_p : w_t = 0:4$ (b) $w_p : w_t = 1:3$ (c) $w_p : w_t = 2:2$ (d) $w_p : w_t = 3:1$ (e) $w_p : w_t = 4:0$

Pen et al., "PMG: Personalized Multimodal Generation with Large Language Models", TheWebConf 2024

Conclusion