User Embedding Model for Personalized Language Prompting

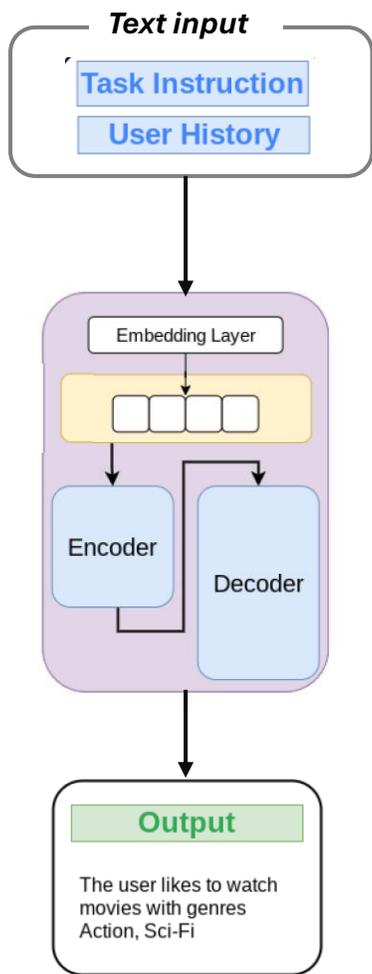Sumanth Doddapaneni, Krishna Sayana*, Ambarish Jash, Sukhdeep Sodhi, Dima Kuzmin
Google Research
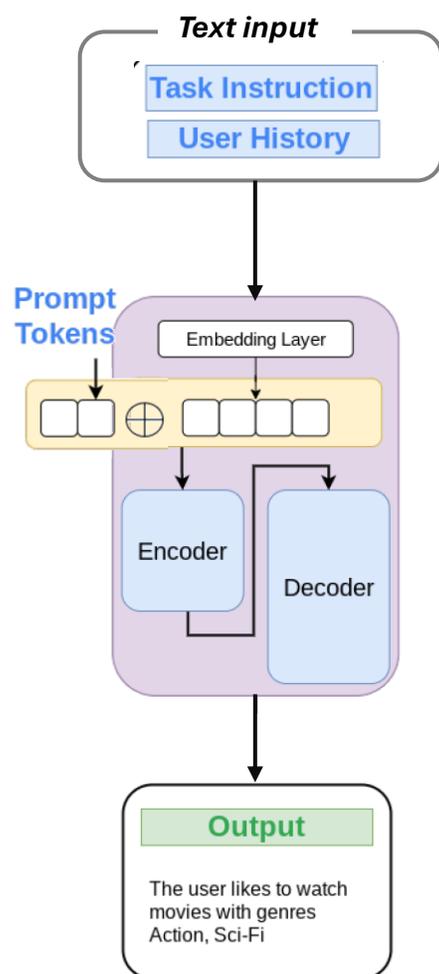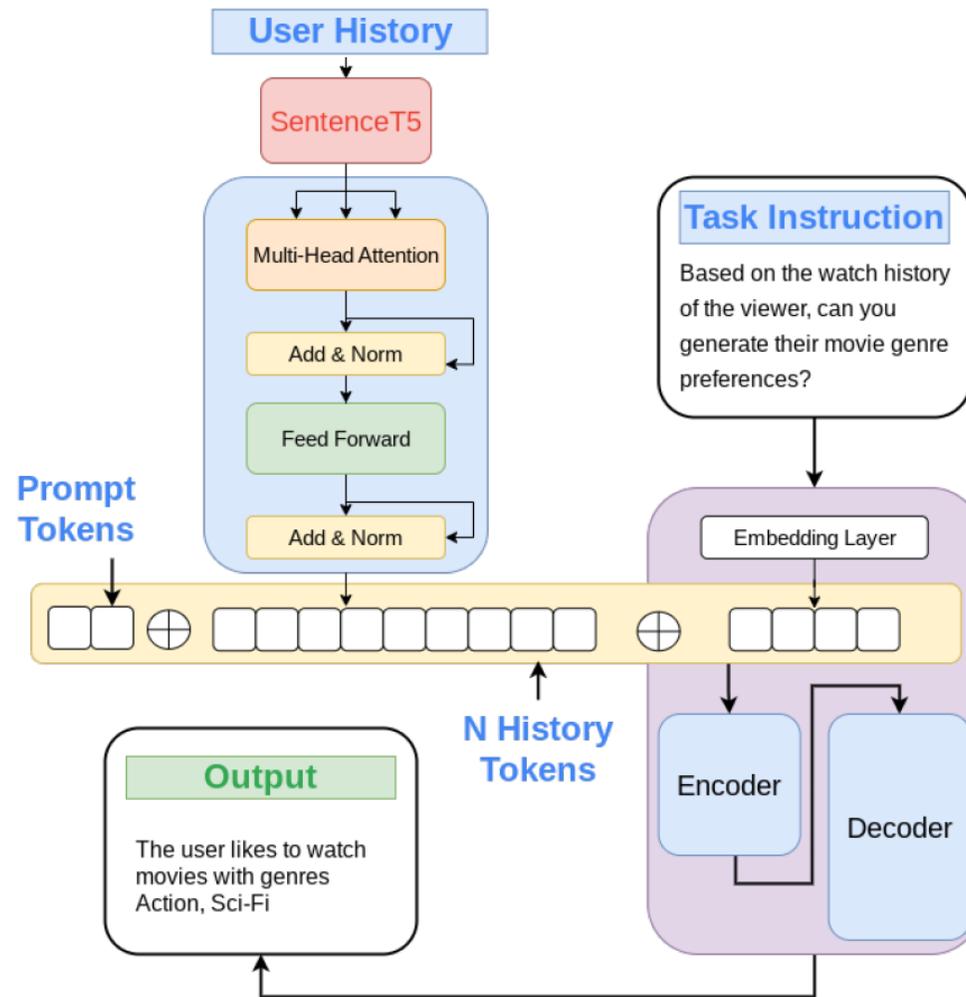
2025.10.26

IDEA LAB.

정유립

Text-based Recommendation (T5)
P5 (Geng et al., 2022)

Soft Prompt Text-based Recommendation
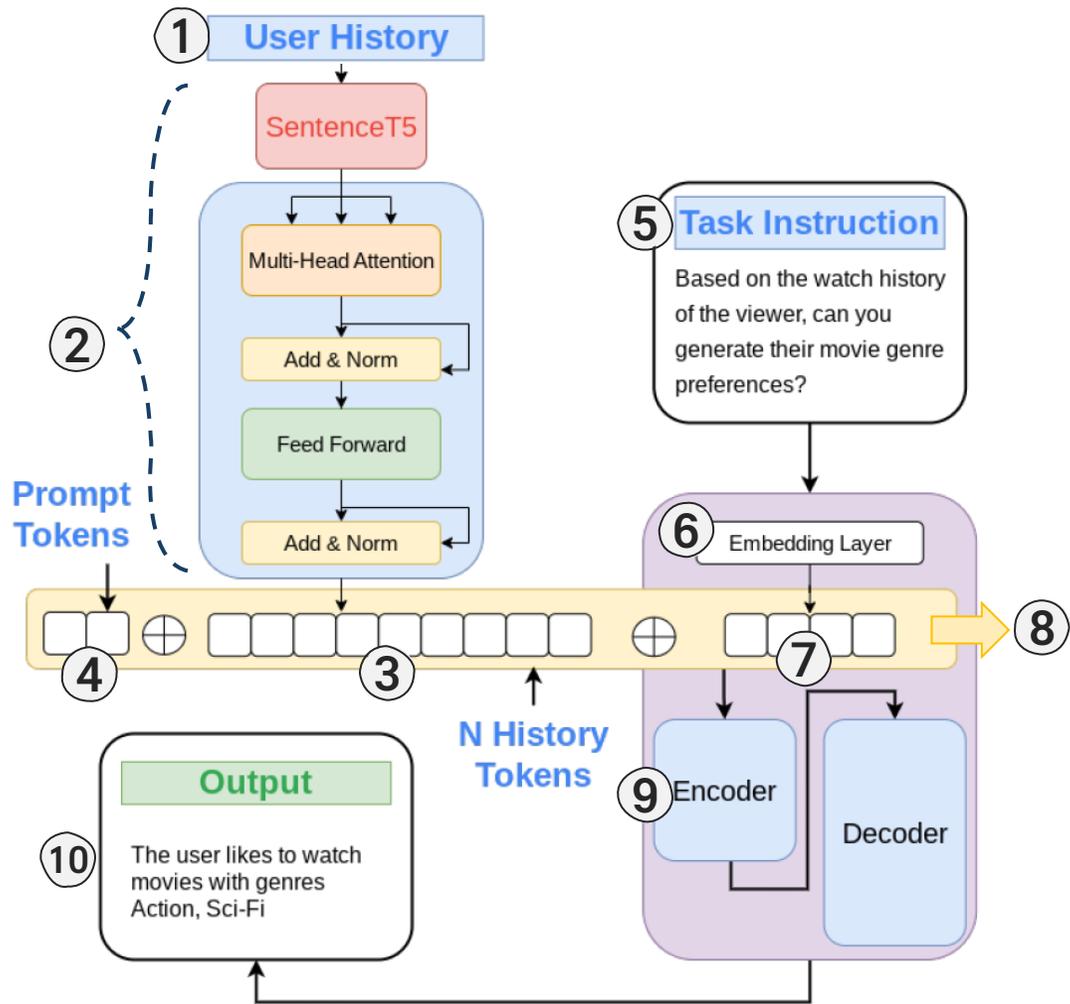(Li et al. 2023)

User embedding Model

Figure 1: Overview of our User History Modeling Approach. The user history's textual features are processed through the User Embedding module and combined with the task prompt and subsequently passed through the language model.

Loss Function

Cross Entropy : $\mathcal{L}(\theta) = -\sum_{t=1}^{|Y|} \log P_\theta \left(y_t \mid y_{<t}, Z\right)$

Notations

#	Notation	Meaning	Type
①	U	user history text	text
②	$UEM(\cdot)$	user embedding module pipeline	function
③	$U_e = UEM(U)$	user personalized prompt embedding	vector matrix
④	K	task-level soft prompt	learnable matrix
⑤	I	task instruction text	text
⑥	$F(\cdot)$	embedding layer (Text → embedding)	function
⑦	$I_e = F(I)$	Task instruction embedding	vector matrix
⑧	$Z = [K; U_e; I_e]$	full input embedding sequence	vector matrix
⑨	$G(\cdot)$	FlanT5 encoder−decoder LM	function
⑩	$\hat{Y} = G(Z)$	Output text	text
-	$y_t , y_{<t}$	target text at time step t & Context at time step t	text

Datasets (MovieLens)

ratings.dat

UserID	MovieID	Rating	Timestamp
1	1193	5	978300760
1	661	3	978302109
1	914	3	978301968
1	3408	4	978300275
...			

movies.dat

MovieID	Title	Genres
1	Toy Story (1995)	Animation|Children's|Comedy
2	Jumanji (1995)	Adventure|Children's|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
...		

input text

title: The Matrix
genre: Action, Sci-Fi
rating: 5
description: A hacker discovers reality is a simulation.
---
title: Titanic
genre: Drama, Romance
rating: 4
description: A love story aboard the doomed Titanic.
...

		BASE	LARGE
Counting Baselines	precision		0.330
	recall		0.273
	f1		0.192
Text Hist. 5	precision	0.276	0.257
	recall	0.287	0.273
	f1	0.273	0.261
Emb. Hist. 5	precision	0.275	0.281
	recall	0.297	0.290
	f1	0.252	0.215
LongT5 50	precision	0.541	0.568
	recall	0.523	0.558
	f1	0.529	0.557
Emb. Hist. 50	precision	0.407	0.400
	recall	0.405	0.399
	f1	0.396	0.381
Emb. Hist. 100	precision	0.416	0.459
	recall	0.413	0.441
	f1	0.404	0.444

Table 1: Model performance using proposed User Embedding Module. Counting Baselines refers to counting the three most frequently occurring genres across the entire user history.
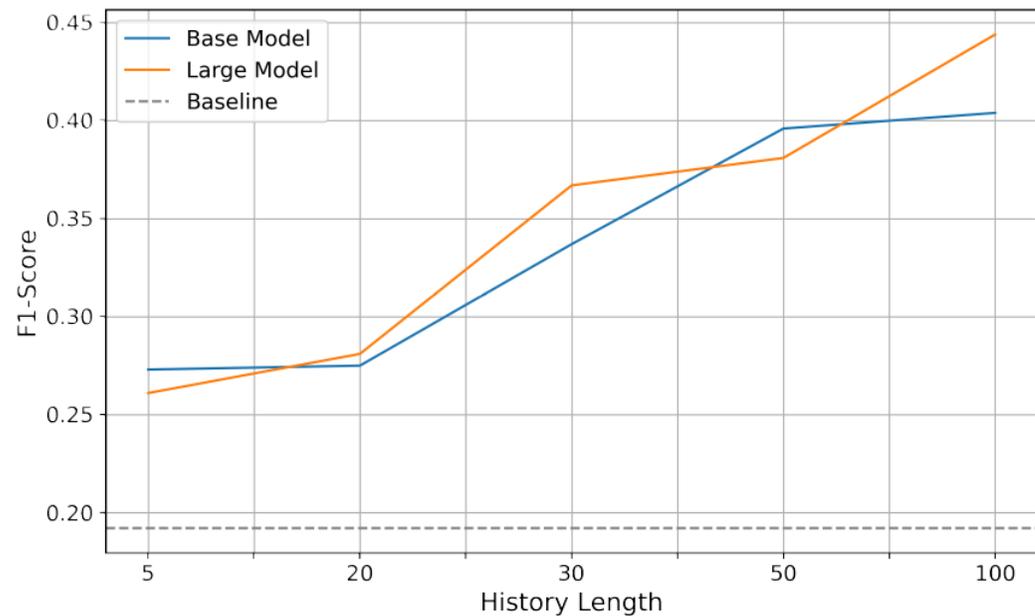

Figure 2: Comparison of model performance with increasing User History.

		BASE	LARGE
$T5_{1.1}$	precision	0.282	0.322
	recall	0.292	0.324
	f1	0.208	0.267
$T5_{LMAdapted}$	precision	0.353	0.398
	recall	0.374	0.397
	f1	0.338	0.378
FlanT5	precision	0.407	0.400
	recall	0.405	0.399
	f1	0.396	0.381

Table 2: Comparison of model performance with various choices of Language Models.

		BASE	LARGE
1 Layer	precision	0.391	0.395
	recall	0.381	0.384
	f1	0.346	0.347
2 Layers	precision	0.399	0.380
	recall	0.392	0.367
	f1	0.384	0.365
3 Layers	precision	0.407	0.400
	recall	0.405	0.399
	f1	0.396	0.381

Table 3: Comparison of model performance with various sizes of User Embedding Module. All the models use the same history size of 50.

		BASE	LARGE
Emb. Hist. 5	precision	0.276	0.257
	recall	0.287	0.273
	f1	0.273	0.261
Emb. Hist. 20	precision	0.319	0.321
	recall	0.328	0.326
	f1	0.275	0.281
Emb. Hist. 30	precision	0.353	0.390
	recall	0.364	0.390
	f1	0.337	0.367
Emb. Hist. 50	precision	0.407	0.400
	recall	0.405	0.399
	f1	0.396	0.381
Emb. Hist. 100	precision	0.416	0.459
	recall	0.413	0.441
	f1	0.404	0.444

Table 4: Comparison of model performance with increasing User History.