南京航空航天大學
Nanjing University of Aeronautics and Astronautics

ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models

Qijiong Liu
liu@qijiong.work
The Hong Kong Polytechnic University
Hong Kong, China

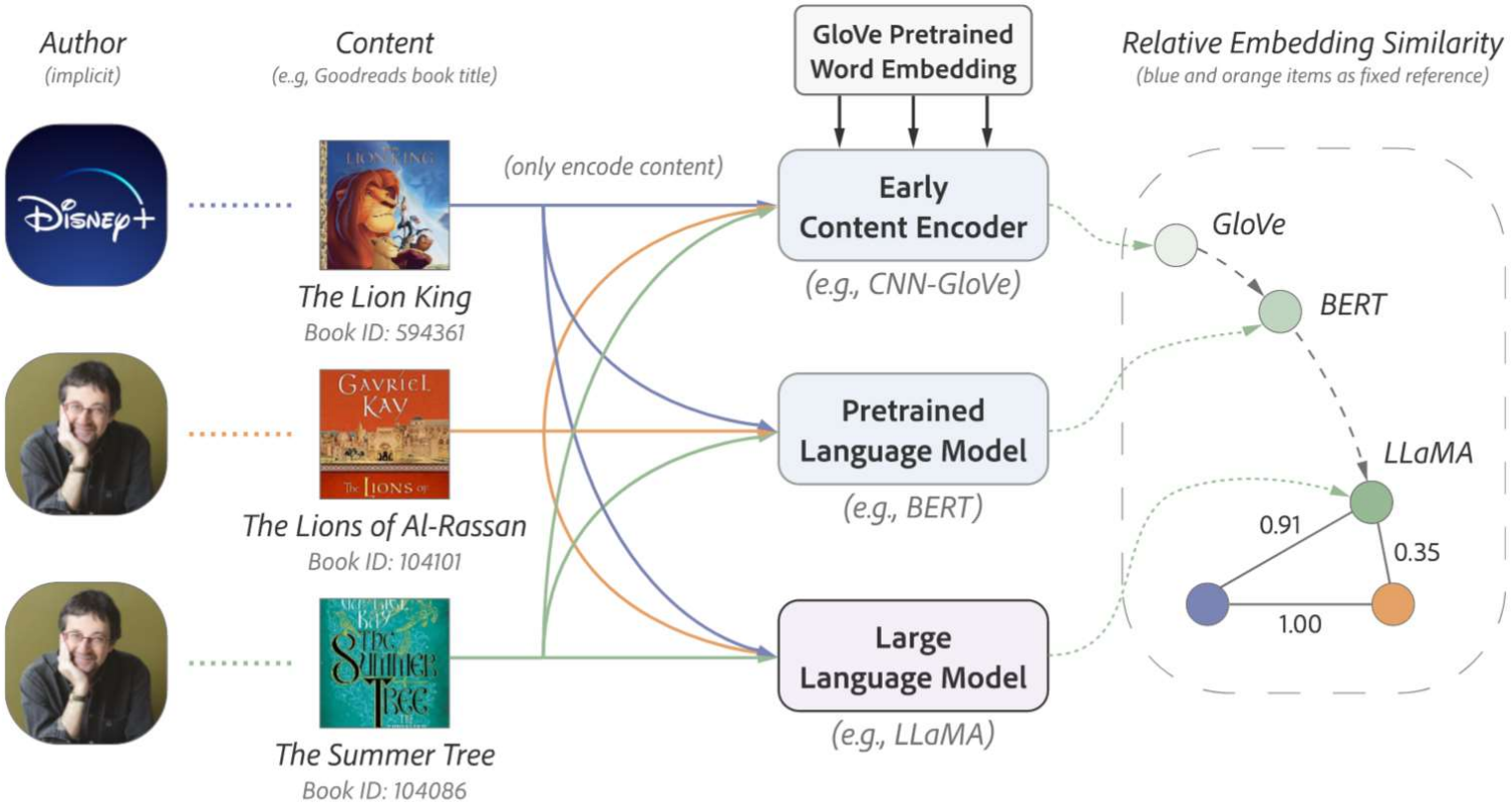
Nuo Chen
pleviumtan@toki.waseda.jp
Waseda University
Tokyo, Japan

Tetsuya Sakai
tetsuyasakai@acm.org
Waseda University
Tokyo, Japan

Xiao-Ming Wu*
xiao-ming.wu@polyu.edu.hk
The Hong Kong Polytechnic University
Hong Kong, China

WSDM 2024

Background

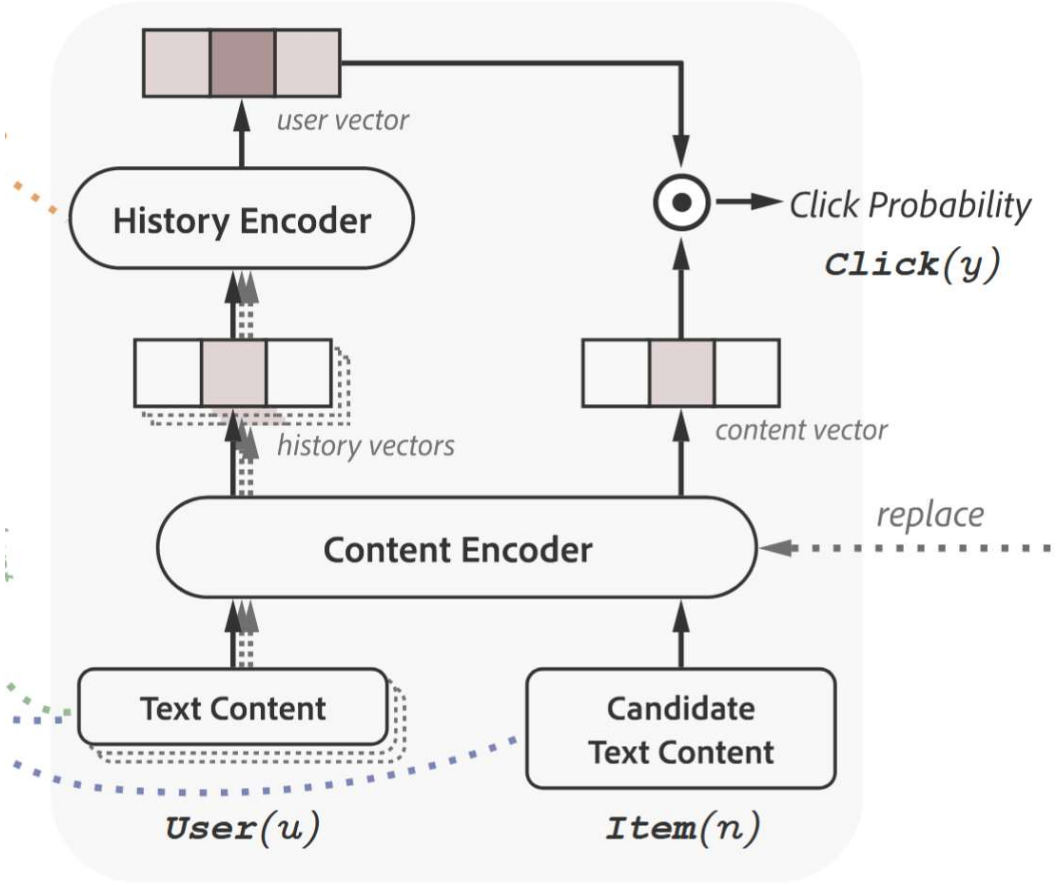


计算余弦相似度: $s_{i,j}$

定义相对距离:

$$d_{i,j} = \frac{1 - s_{i,j}}{1 - s_{\text{blue,orange}}}$$

Content-based Recommendation



Content Encoder

内容编码器的任务是将每个内容的多个特征进行编码，并将它们整合成一个统一的 d-维向量v_n，便于后续计算和处理。

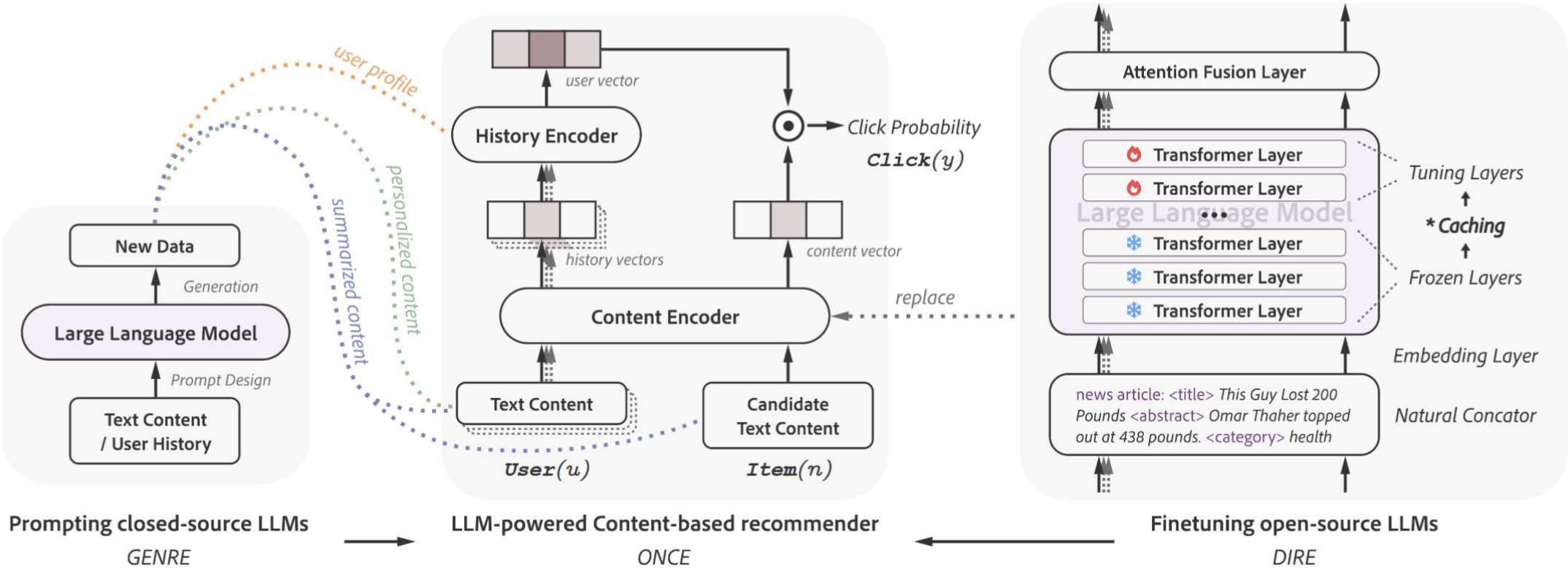
History Encoder

历史编码器的作用是根据用户的浏览历史生成一个统一的d-维用户向量v_u。用户的浏览历史是一个序列，由多个内容向量组成，历史编码器会根据这些内容向量生成一个表示用户兴趣的向量。

Interaction Module

交互模块的目标是从多个候选内容向量中，找出最符合用户兴趣的正样本内容。交互模块的任务就是通过某种方式来识别出与用户兴趣最匹配的正样本。

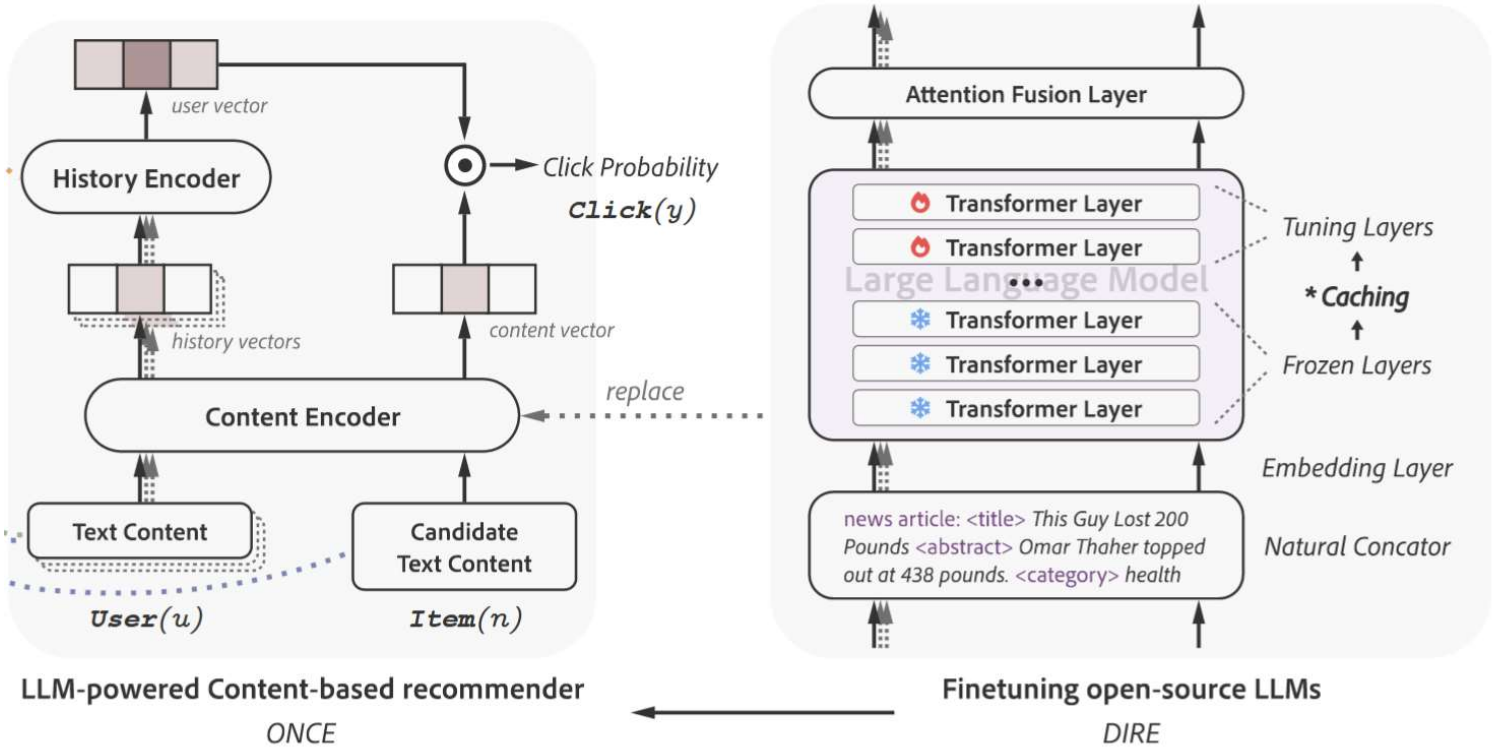
ONCE Open- and Closed-source LLMs



[CLS] AI技术的未来 [SEP] 人工智能将在各个领域改变世界 [SEP] 科技 [SEP]

DIRE

南京航空航天大學 Nanjing University of Aeronautics and Astronautics



Embedding Layer

$$\mathbf{E}^{\mathbf{0}} = EmbeddingLayer\,(\mathbf{s}) \in \mathbb{R}^{l \times d^n}$$

Transformer Decoder

$$\mathbf{E}^{\mathbf{i}} = TransformerLayer\left(\mathbf{E}^{\mathbf{i-1}}\right) \in \mathbb{R}^{l \times d^n}$$

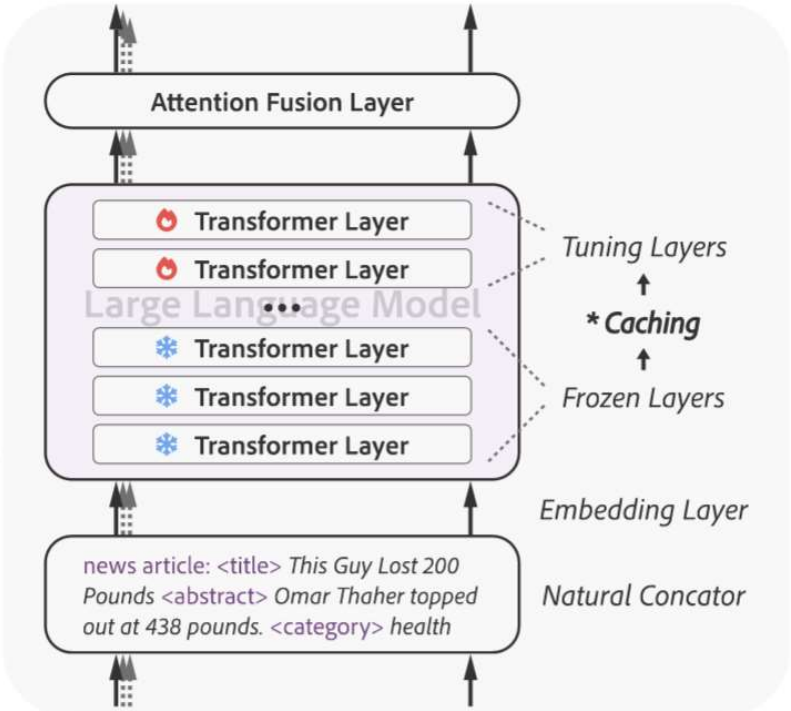
$$i \in \{1, \ldots, H\},$$

Attention Fusion Layer

$$\mathbf{Z} = \mathbf{E}^{\mathbf{i}}\mathbf{W} + \mathbf{b} \in \mathbb{R}^{l \times d},$$

$$\mathbf{z} = Attention\,(\mathbf{Z}) \in \mathbb{R}^{d},$$

Finetuning Strategy



Finetuning open-source LLMs

DIRE

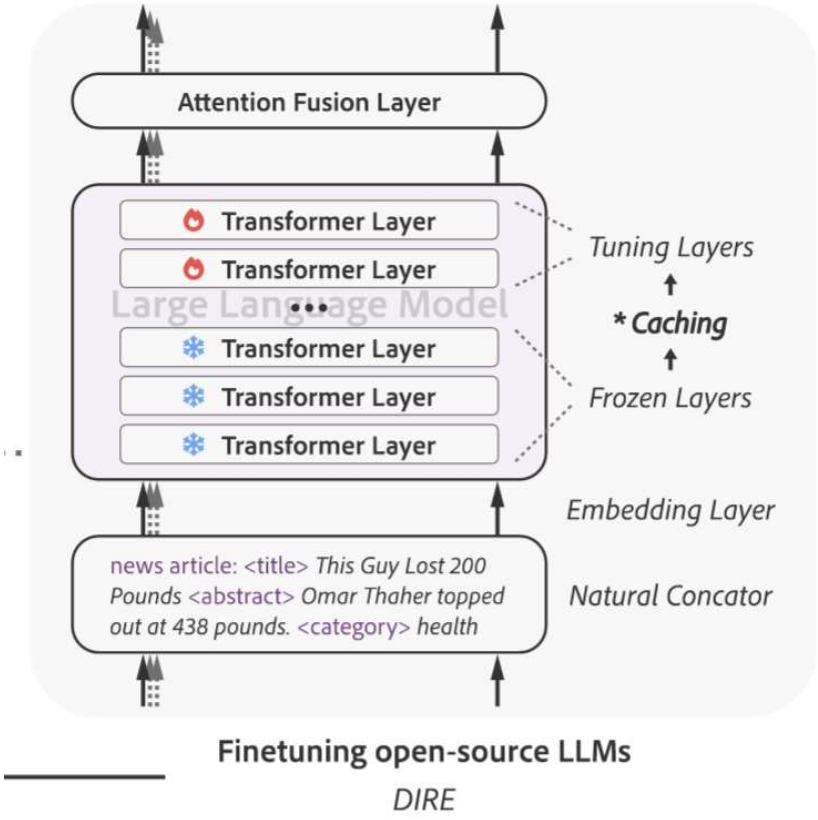
Partial Freezing

冻结低层，微调高层：大型语言模型（LLM）通常由多个Transformer 层组成，这些层的参数非常庞大，训练时的计算量也极其高。LLM的**低层**通常会学习到一些比较通用的特征，而这些特征对于特定任务的微调帮助不大。
所以，作者决定“冻结”这些低层的参数，也就是说，在训练时不更新这些层的权重，**只微调高层**。这样可以大大减少计算负担。

Caching

具体来说，在开始微调之前，他们先预计算并存储所有内容的**隐藏状态**。这些隐藏状态来自于被冻结的低层，意味着每个内容的表示已经计算完成，可以在后续的训练中直接使用，而不需要每次训练时都重新计算。这使得训练过程中的计算量大大减少。

Finetuning Strategy



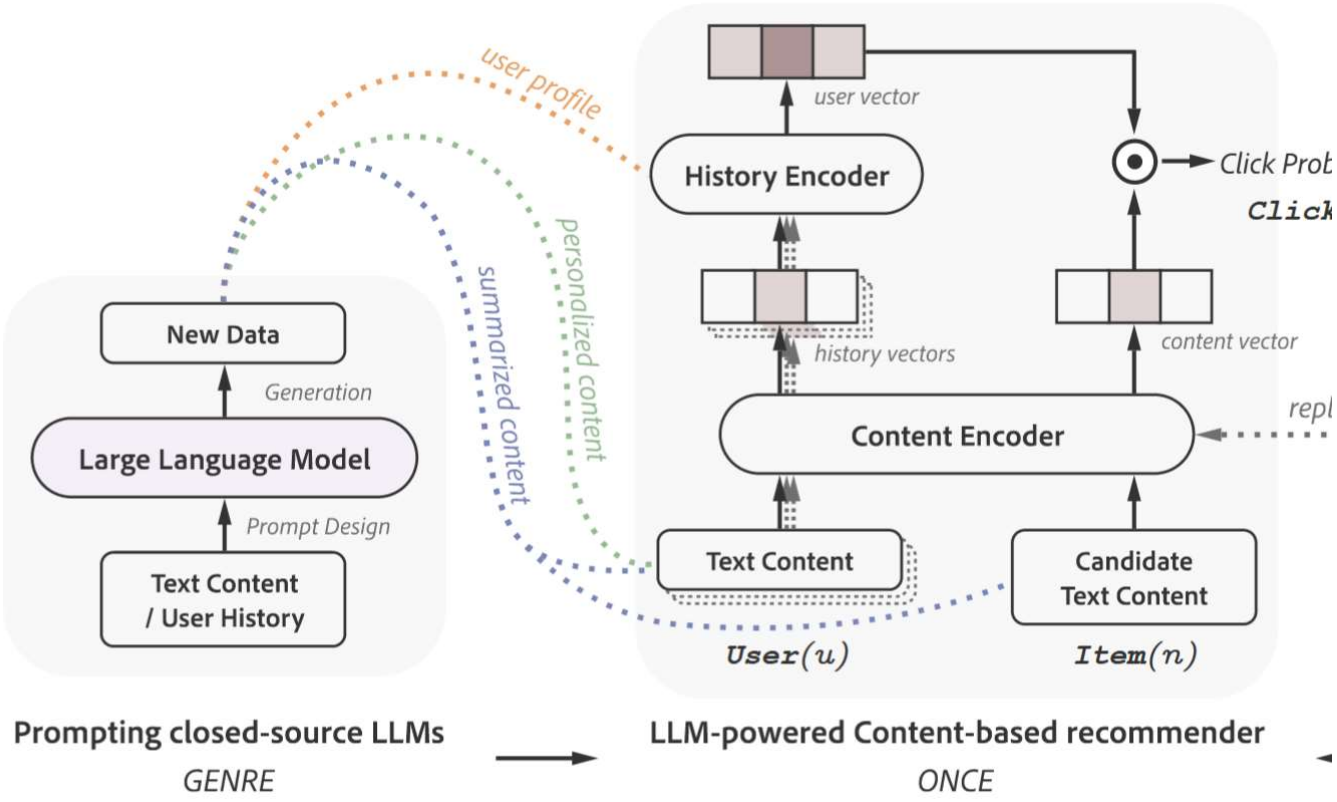
Finetuning open-source LLMs

DIRE

LoRA

它通过分解原始的高维权重矩阵为两个低秩矩阵，这样减少了模型训练时需要调整的参数数量，**大大降低了训练的复杂性和计算开销**。
作者将LoRA应用到被冻结的Transformer层上，这些层通常是模型中参数最多的部分。通过LoRA的优化，可以**极大地减少需要微调的参数数量**，提高训练效率。

GENRE



Content Summarizer

大语言模型（LLMs）能够有效地对文本内容进行总结，从而生成更加精简和信息量丰富的内容表示。

User Profiler

分析用户的兴趣、偏好、地理位置等特征，从而为推荐系统提供**更精准**的用户表示。

Personalized Content Generator

通过少量用户交互数据为用户生成个性化的内容，用于缓解**冷启动**的问题

Content Summarizer

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

prompt

> *You are asked to act as a news title enhancer. I will provide you a piece of news, with its original title, category, subcategory, and abstract (if exists). The news format is as below:*
>
> *[title] {title}*
> *[abstract] {abstract}*
> *[category] {category}*
> *[subcategory] {subcategory}*
>
> *where {title}, {abstract}, {category}, and {subcategory} will be filled with content. You can only response a rephrased news title which should be clear, complete, objective and neutral. You can expand the title according to the above requirements. You are not allowed to response any other words for any explanation. Your response format should be:*
>
> *[newtitle] {newtitle}*
>
> *where {newtitle} should be filled with the enhanced title. Now, your role of news title enhancer formally begins. Any other information should not disturb your role.*

content

> *[title] Here's Exactly When To Cook Every Dish For Thanksgiving Dinner*
> *[abstract] Time out turkey day down to the minute.*
> *[category] foodanddrink*
> *[subcategory] tipsandtricks*

output

> *[newtitle] Perfectly Timed Thanksgiving Dinner: A Comprehensive Guide to Cooking Every Dish with Minute-by-Minute precision*

User Profiler

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

prompt

You are asked to describe user interest based on his/her browsed news list, the format of which is as below:

(1) {news title}
...
(n) {news title}

You can only response the user interests with the following format to describe the [topics] and [regions] of the user's interest

[topics]
- topic1
- topic2
...
[region] (optional)
- region1
- region2
...

Only [topics] and [region] can be appeared in your response. If you think region are hard to predict, leave it blank. Your response topic/region list should be ordered, that the first several options should be most related to the user's interest. You are not allowed to response any other words for any explanation or note. Now, the task formally begins. Any other information should not disturb you.

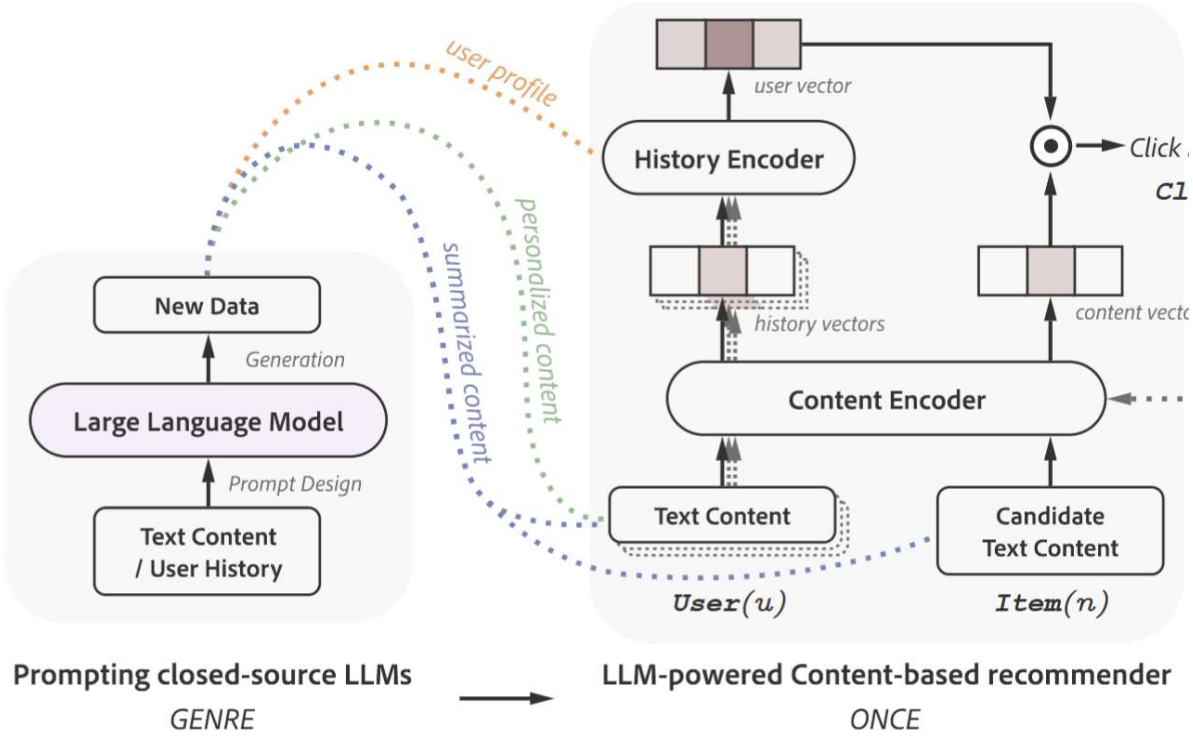
content

(1) Judge: Brad Pitt, others can be sued over New Orleans homes
(2) Solomons vetoes Chinese 'lease' on Pacific island
...
(30) Republicans urge Bevin to provide proof of election fraud or concede

output

[topics]
- legal issues
- politics
- crime
...
[region]
Null

User Profiler



生成兴趣向量v_i

$$\mathbf{v}_i = \left[\mathrm{POOL}\left(\mathbf{E}_{\text{topics}}\right); \mathrm{POOL}\left(\mathbf{E}_{\text{regions}}\right)\right] \in \mathbb{R}^{2\times d},$$

兴趣感知的用户向量v_{iu}

$$\mathbf{v}_{iu} = \mathrm{MLP}\left([\mathbf{v}_u; \mathbf{v}_i]\right) \in \mathbb{R}^{d},$$

Personalized Content Generator

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

prompt

You are asked to capture user's interest based on his/her browsing history, and generate a piece of news that he/she may be interested. The format of history is as below:

(1) (the category of the first news) the title of the first news
...
(n) (the category of the n-th news) the title of the n-th news

You can only generate a piece of news (only one) in the following json format:

{"title": <title>, "abstract": <news abstract>, "category": <news category>}

where <news category> is limited to the following options:

"title", "abstract", and "category" should be the only keys in the json dict. The news should be diverse, that is not too similar with the original provided news list. You are not allowed to response any other words for any explanation or note. JUST GIVE ME JSON-FORMAT NEWS. Now, the task formally begins. Any other information should not disturb you.

content

(1) (entertainment) Meghan King Edmonds, Jim Edmonds' Nanny Denies Cheating Allegations
(2) (tv) Property Brothers' J.D. Scott Marries Annalee Belle in Vintage Theatre-Themed Wedding
...
(4) (tv) Jim Edmonds Calls Police on Meghan King Edmonds Out of "Concern" For Their Kids

output

```
{
  "title": "Top 10 Best Places to Travel in 2020",
  "abstract": "Explore the world's most exciting destinations in the year 2020.",
  "category": "travel"
}
```

```
{
  "title": "10 Delicious Fall-Inspired Recipes To Try This Season",
  "abstract": "Celebrate the arrival of fall with these scrumptious recipes that will warm you up on chilly evenings.",
  "category": "foodanddrink"
}
```

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Datasets：

MIND，Goodreads

Recommendation Models：

NAML， NRMS， Fastformer

Evaluation Metrics：

AUC ， MRR， nDCG

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

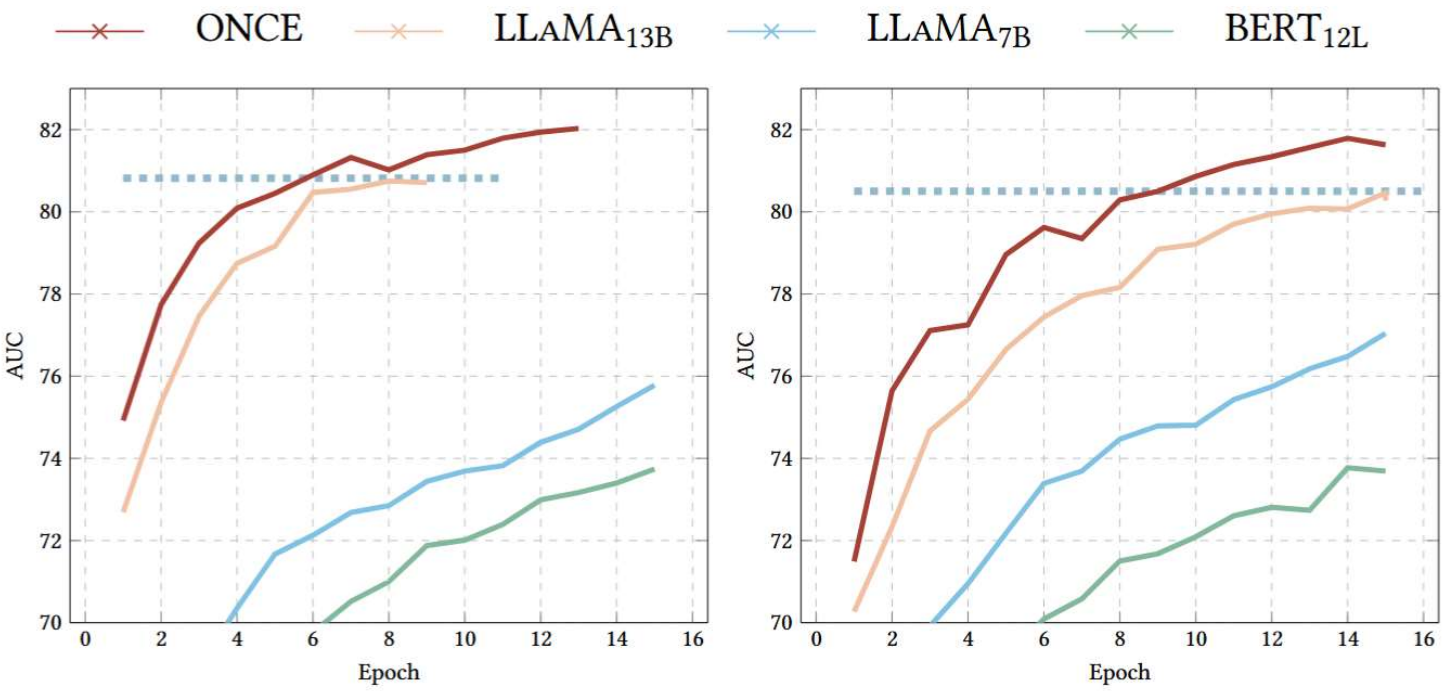
Dataset	MIND	Goodreads	Dataset	MIND	Goodreads
Original			*Content Summarizer (CS)*		
# content	65,238	16,833	**tokens/title**	+3.17	-
tokens/title	13.56	6.10	**tokens/desc**	-	29.28
# users	94,057	23,089	*User Profiler (UP)*		
# new user	20,110	2,306	**topics/user**	4.82	4.55
content/user	14.98	7.81	**regions/user**	0.29	-
content/user$_n$	3.19	3.03	*Personalized Content Generator (CG)*		
# pos	347,727	273,888	**#content**	+40,220	+4,612
# neg	8,236,715	485,233	**content/user$_n$**	+2.00	+2.00

Experiments

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

		NAML (2019a)				NRMS (2019c)				Fastformer (2021b)				MINER (2022)			
		AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10	AUC	MRR	N@5	N@10
MIND dataset																	
	Original	61.75	30.60	31.35	37.85	61.71	30.20	30.98	37.42	62.26	31.14	31.90	38.32	63.88	32.19	33.04	39.45
DIRE	**BERT$_{12L}$** [46]	65.32	33.16	34.29	40.35	64.08	31.24	32.35	38.66	65.48	32.47	33.41	39.75	65.82	32.77	34.02	40.19
	LLaMA$_{7B}$ *(Ours)*	68.34	35.80	37.60	43.48	68.50	36.21	38.11	43.91	68.55	36.59	38.38	44.06	68.70	36.58	38.49	44.18
	LLaMA$_{13B}$ *(Ours)*	68.23	35.99	37.93	43.77	68.45	36.15	38.02	43.88	68.51	36.37	38.20	44.02	68.59	36.46	38.38	44.05
GENRE	**CS** *(Ours)*	63.73	31.83	32.94	39.24	63.85	31.57	32.35	38.80	64.73	32.81	33.68	40.06	65.71	33.59	34.90	40.96
	UP *(Ours)*	62.19	30.90	31.78	38.26	61.90	30.60	31.54	37.66	63.40	31.94	32.76	39.15	64.45	32.09	33.14	39.54
	CG *(Ours)*	62.93	30.83	32.10	38.34	63.04	31.00	31.84	38.22	64.69	32.28	33.31	39.76	64.21	32.30	33.57	39.91
	UP→CG *(Ours)*	63.61	31.58	32.63	39.07	62.95	32.00	32.80	39.00	64.82	32.44	33.51	39.93	64.73	33.09	34.10	40.32
	ALL *(Ours)*	63.88	32.17	33.14	39.37	63.71	32.14	33.11	39.43	66.70	34.20	35.81	41.78	66.46	34.20	35.47	41.48
	ONCE *(ours)*	**68.62**	**36.50**	**38.31**	**44.05**	**68.74**	**36.66**	**38.60**	**44.37**	**68.83**	**36.68**	**38.56**	**44.35**	**68.92**	**36.74**	**38.72**	**44.48**
	Improvement (%) over Original	**11.13%**	**19.28%**	**22.20%**	**16.38%**	**11.39%**	**21.39%**	**24.60%**	**18.57%**	**10.55%**	**17.79%**	**20.88%**	**15.74%**	**7.89%**	**14.13%**	**17.19%**	**12.75%**
	Improvement (%) over BERT$_{12L}$	**5.05%**	**10.07%**	**11.72%**	**9.17%**	**7.27%**	**17.35%**	**19.32%**	**14.77%**	**5.12%**	**12.97%**	**15.41%**	**11.57%**	**4.71%**	**12.11%**	**13.82%**	**10.67%**

Experiments



(a) NRMS (b) Fastformer

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Thanks