Personalized Prompt Learning for Explainable Recommendation

Lei Li[1], Yongfeng Zhang[2], Li Chen[1]

[1] Hong Kong Baptist University, [2] Rutgers University

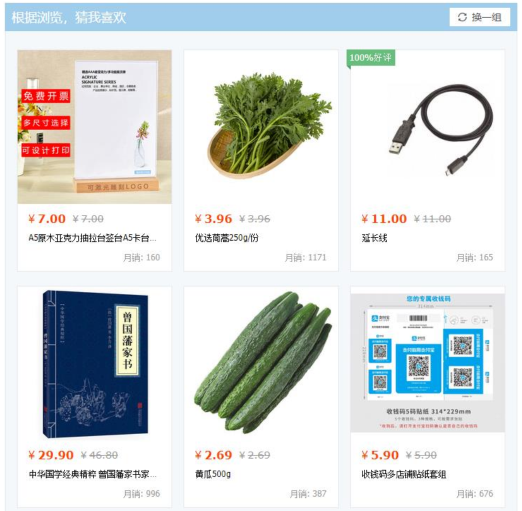
csleili@comp.hkbu.edu.hk

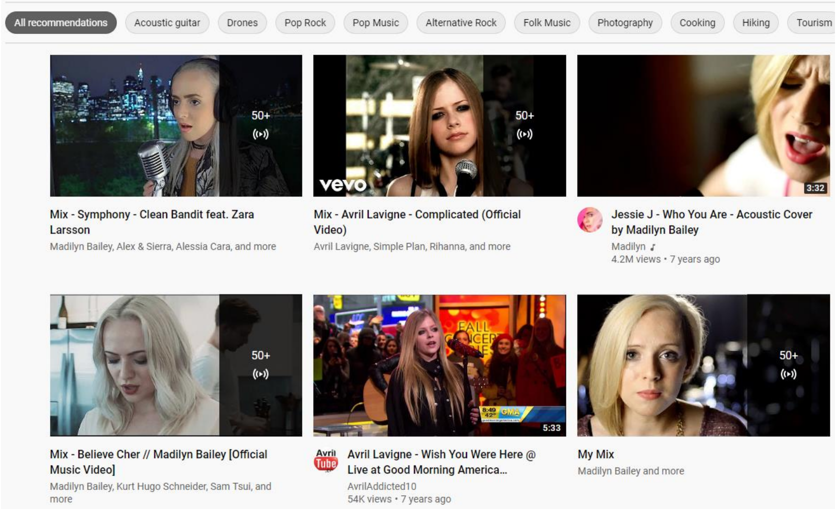
Dec. 11, 2024

The 2nd International ACM SIGIR Conference on Information Retrieval in the Asia Pacific (SIGIR-AP'24)

Recommendations Everywhere

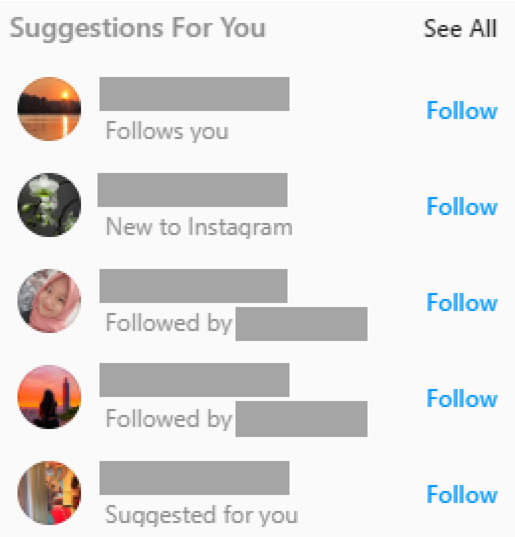
E-commerce
(taobao.com)





Video
(youtube.com)

Social Media
(instagram.com)

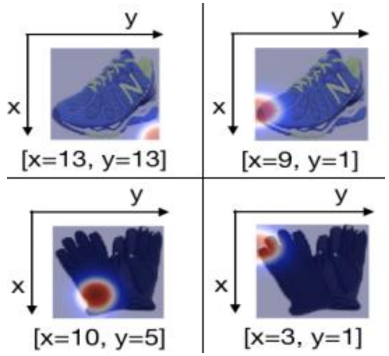


Movie
(movie.douban.com)

Explainable Recommendation

You might be interested in [feature], on which this product performs well.
You might be interested in [feature], on which this product performs poorly.

- Given a user-item pair, provide an **explanation** to justify why the item is recommended to the user
  - Pre-defined template [1]
  - Image visualization [2]
  - Natural language sentence in this work
    - *"the style of the jacket is fashionable"*



[1] Zhang, Yongfeng, et al. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis." SIGIR'14.
[2] Chen, Xu, et al. "Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network: Towards Visually Explainable Recommendation." SIGIR'19.

Explanatory Goals

- **Trust**: increase users' confidence in the system
- **Effectiveness**: help users make good decisions
- **Persuasiveness**: convince users to try or buy
- **Efficiency**: help users make decisions faster
- **Satisfaction**: increase the ease of use or enjoyment

User-centric

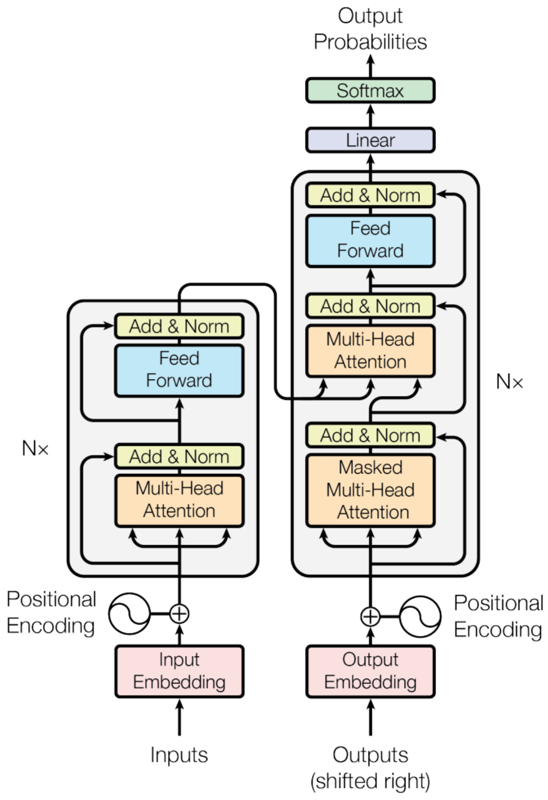
- **Transparency**: explain how the system works
- **Scrutability**: allow users to tell the system it is wrong

System-centric

[1] Tintarev, Nava, and Judith Masthoff. "Explaining recommendations: Design and evaluation." Recommender systems handbook. 2015.

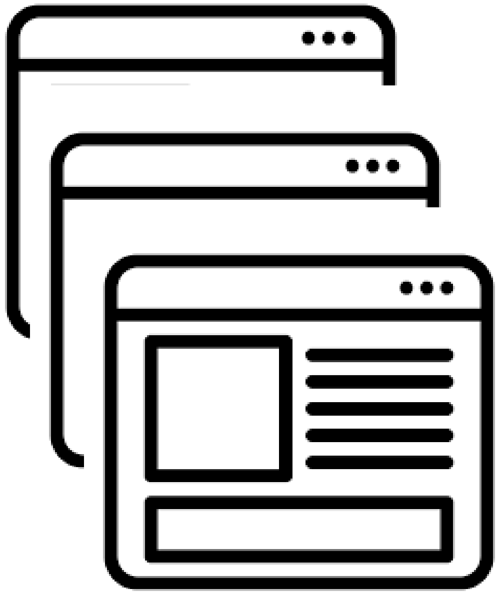
Pre-trained Language Models (1)

Transformer [1] with huge amount of parameters



Large corpora from Internet

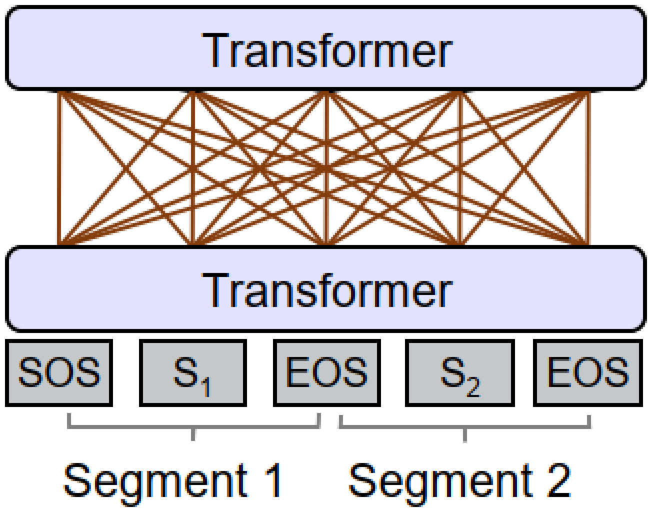
- News articles
- Wikipedia webpages
- Blogs
- ……



[1] Vaswani, Ashish, et al. "Attention is all you need." NIPS'17.

Pre-trained Language Models (2)

BERT [1]

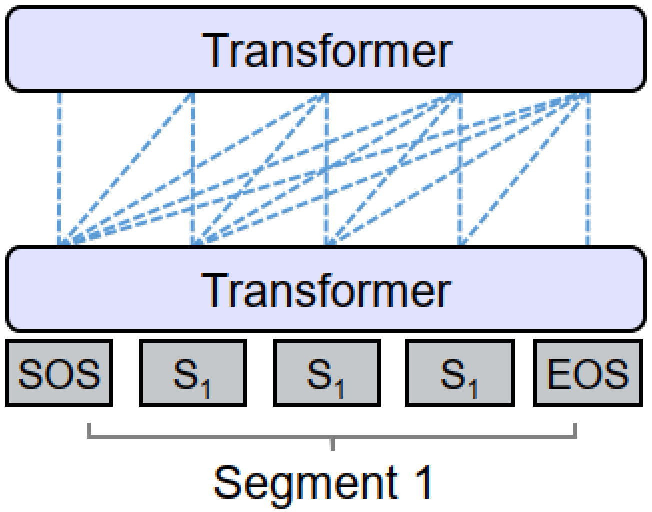


Bidirectional model for classification

VS.

Image credit to [3]

GPT [2]



Unidirectional model for generation

Difficult to do customized modifications

[1] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." NAACL'19.
[2] Radford, Alec, et al. "Improving language understanding by generative pre-training." 2018.
[3] Dong, Li, et al. "Unified language model pre-training for natural language understanding and generation." NeurIPS'19.

Prompt Learning

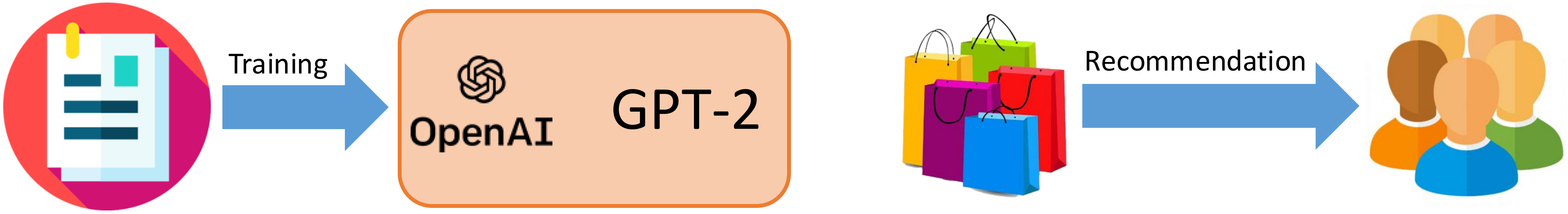
- Adapt different tasks to pre-trained language models
- Directly model text probability
- But require human labor to design templates and answer set

Task	Input (X)	Template	Output (Y)
Sentiment Classification	I love this book.	X The book is Y *prompt*	great boring ...
Text Summarization	The Omicron ...	X TL;DR: Y	COVID-19 ... Pandemic ... ...
Machine Translation	Elle m'a apprivoisé.[3]	French: X English: Y	She tamed ... The flower ... ...

[1] Liu, Pengfei, et al. "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing." ACM Computing Surveys. 2023.

Prompt Learning for Recommender Systems

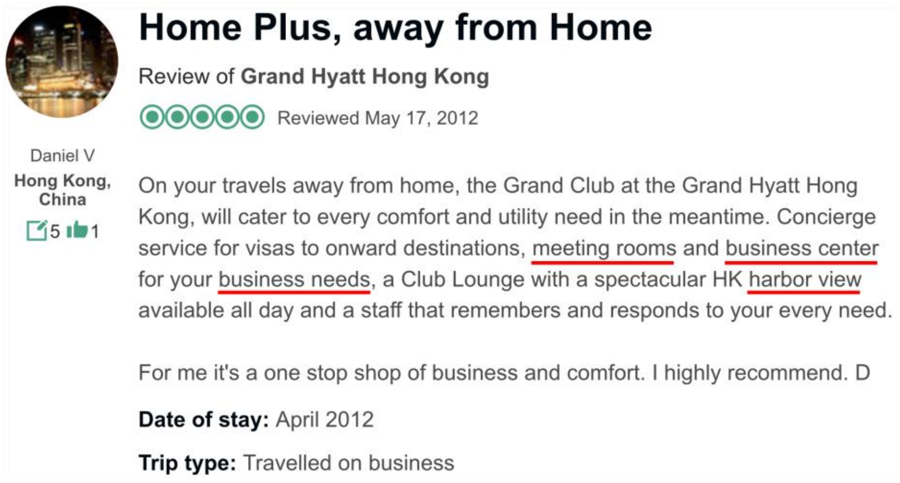
- Language models (LM) were trained with textual data
- User/item IDs in recommender systems are not text

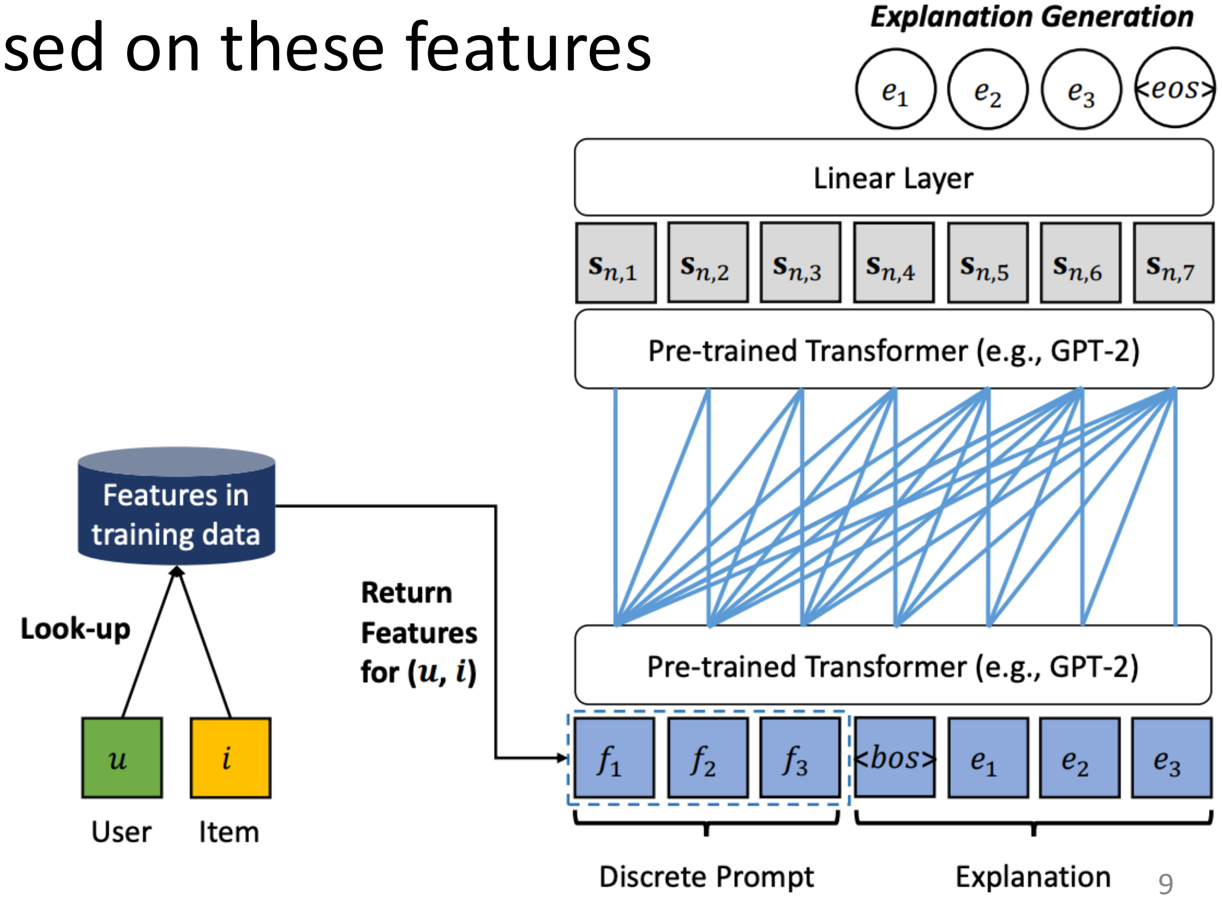


How to incorporate IDs into LM?

Discrete Prompt Learning

- Use item features mined from reviews as an alternative of IDs
- Perform seq2seq generation based on these features





Why Omit the Template?

- A study [1] showed that templates are not always useful for pre-trained language models.
- Our key focus is on automatic explanation generation rather than template design.
- The pre-trained language model in this work is task-specific.

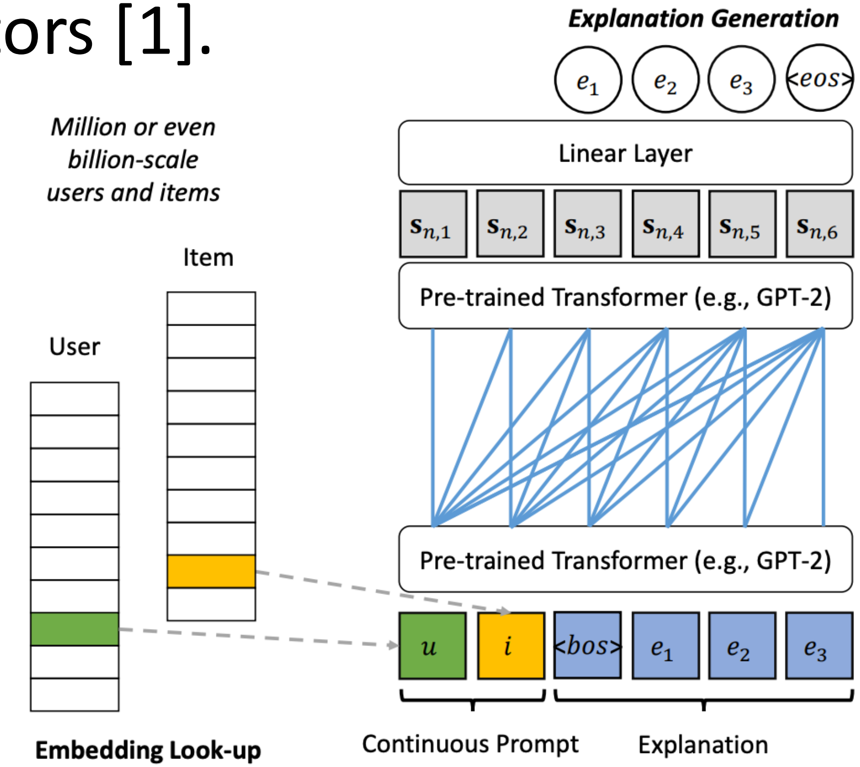
Task	Input (X)	Template	Output (Y)
Explanation Generation	room location ...	X (Explain the recommendation:) Y	The room ... The breakfast ...
	X1: user123abc X2: item456def	X1 X2 (Explain the recommendation:) Y	The location ... ...

[1] Li, Xiang Lisa, and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation." ACL'21.

Continuous (Soft) Prompt Learning

- The conversion from IDs to item features loses information.
  - E.g., Jerry vs. Cheese
- Prompts can be human-incomprehensible vectors [1].

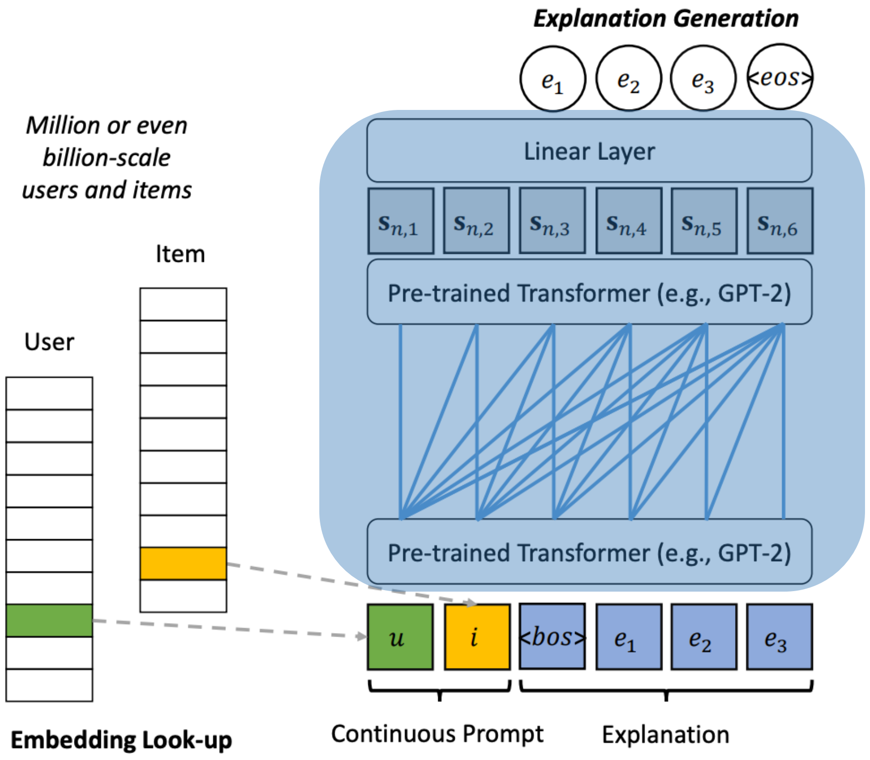
- Solution
  - Associate each ID with a vector
  - Pass vectors of target user and item to the model
  - Perform auto-regressive generation



[1] Li, Xiang Lisa, and Percy Liang. "Prefix-tuning: Optimizing continuous prompts for generation." ACL'21.

Sequential Tuning

- Semantic gap: randomly initialized ID vectors vs. pre-trained LM
- Randomly initialized vectors could not be well optimized with SGD [1].



LM parameters fixed

Fixed-LM Prompt Tuning

Prompt+LM Fine-tuning

[1] Allen-Zhu, Zeyuan, et al. "A convergence theory for deep learning via over-parameterization." ICML’19.

Datasets

- Yelp
  - Restaurant
- Amazon
  - Movies & TV
- TripAdvisor
  - Hotel
- An explanation is a review sentence that contains item feature(s).

	Yelp	Amazon	TripAdvisor
#users	27,147	7,506	9,765
#items	20,266	7,360	6,280
#records	1,293,247	441,783	320,023
#features	7,340	5,399	5,069
#records / user	47.64	58.86	32.77
#records / item	63.81	60.02	50.96
#words / exp	12.32	14.14	13.01

* **exp** denotes **explanation**.

[1] Li, Lei, et al. "Generate neural template explanations for recommendation." CIKM'20.

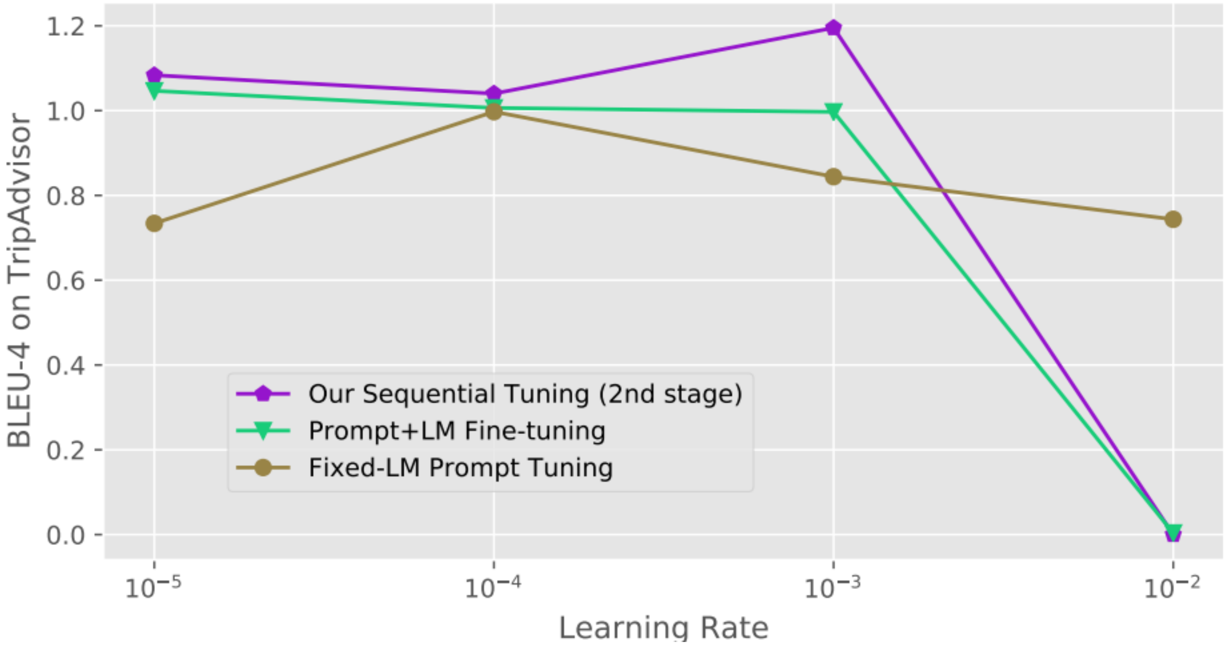
Evaluation Metrics

- Text quality: not equal to explainability
  - BLEU [1]
  - ROUGE [2]
- Explainability from the angle of item features [3]
  - Unique Sentence Ratio (USR)
  - Feature Matching Ratio (FMR)
  - Feature Coverage Ratio (FCR)
  - Feature Diversity (DIV)

[1] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL'02.
[2] Lin, Chin-Yew. "ROUGE: A Package for Automatic Evaluation of Summaries." ACL'04 Workshop.
[3] Li, Lei, et al. "Generate neural template explanations for recommendation." CIKM'20.

Effect of Sequential Tuning

- Sequential tuning is more effective than prompt+LM tuning and prompt tuning.
- Sequential tuning and prompt+LM tuning are more sensitive to large learning rates.



Quantitative Results

- Our approaches outperform baselines.
- Continuous prompts are more effective than discrete prompts.

	Explainability				Text Quality							
	FMR↑	FCR↑	DIV↓	USR↑	B1↑	B4↑	R1-P↑	R1-R↑	R1-F↑	R2-P↑	R2-R↑	R2-F↑
	Yelp											
ACMLM	**0.05**	**0.31**	**0.95**	**0.95**	7.01	0.24	7.89	7.54	6.82	0.44	0.48	0.39
PEPLER-D	**0.05**	0.24	1.53	0.13	**9.17****	**0.40****	**15.67****	**10.47****	**11.73****	**1.09****	**0.78****	**0.83****
NRT	0.06	0.12	1.67	0.20	10.92	0.60	16.73	11.91	12.89	1.63	1.21	1.26
Att2Seq	0.05	0.05	2.25	0.05	10.25	0.54	17.13	11.44	12.72	1.49	1.13	1.16
PETER	**0.08**	0.15	1.62	0.15	10.74	0.63	16.18	11.90	12.63	1.60	1.32	1.28
PEPLER	**0.08****	**0.30****	**1.52**	**0.35****	**11.23**	**0.73****	**17.51**	**12.55***	**13.53****	**1.86***	**1.42**	**1.46****
	Amazon											
ACMLM	**0.10**	**0.31**	2.07	**0.96**	9.52	0.22	11.65	10.39	9.69	0.71	0.81	0.64
PEPLER-D	0.08	0.19	**1.85***	0.15	**10.94****	**0.49****	**16.31****	**11.80****	**12.80****	**1.43****	**1.13****	**1.16****
NRT	0.10	0.04	2.71	0.09	12.06	0.69	17.17	13.15	13.83	1.94	1.68	1.64
Att2Seq	0.09	0.04	2.64	0.05	12.07	0.73	18.35	12.86	14.14	2.01	1.56	1.61
PETER	0.09	0.09	2.16	0.20	11.75	0.89	16.51	13.10	13.55	1.96	1.76	1.68
PEPLER	**0.11**	**0.27****	**2.06**	**0.38****	**13.19***	**1.05****	**18.51**	**14.16**	**14.87**	**2.36***	**1.88**	**1.91****
	TripAdvisor											
ACMLM	**0.07**	**0.41**	**0.78**	**0.94**	3.45	0.02	4.86	3.82	3.72	0.18	0.20	0.16
PEPLER-D	0.05	0.22	2.69	0.08	**14.61****	**0.87****	**18.07****	**14.83****	**15.32****	**1.76****	**1.66****	**1.58****
NRT	0.05	0.02	6.07	0.00	13.76	0.80	19.01	14.57	15.58	2.10	1.59	1.68
Att2Seq	0.06	0.05	4.74	0.02	15.20	0.96	18.74	**16.42**	**16.38**	2.42	**2.32**	**2.19**
PETER	**0.07**	0.09	3.62	0.05	15.13	1.00	18.30	16.15	16.00	2.24	2.23	2.06
PEPLER	**0.07***	**0.21****	**2.71****	**0.24****	**15.49**	**1.09**	**19.48**	15.67	16.24	**2.48**	2.21	2.16

Case Study

	Ground-truth	the swimming **pool** is fantastic
random words	ACMLM	swimming **pool** swimming **pools pool** strip beach area
identical	NRT	the hotel is located in a great location
	Att2Seq	the hotel is located in the heart of the city and the main shopping area is also within walking distance
miss the point	PETER	the hotel is located in the heart of the city and the harbour
	PEPLER-D	the room was very nice and the bed was very comfortable
	PEPLER	the **pool** is amazing and the **pool** is very relaxing
	Ground-truth	this is one of the finest **hotels** in all of Europe
	ACMLM	swimming pool area pool ja ##cu ##zzi pool city area gym building pool area spa gym pool area
	NRT	the **hotel** is located in a great location
	Att2Seq	the **hotel** is located in the heart of the city and the main shopping area is also within walking distance
	PETER	the **hotel** is in a great location
	PEPLER-D	the **hotel** is a short walk from the old town
good-quality	**PEPLER**	the **hotel** is located in the heart of the city and is very well maintained

NLG4RS Eco-system

github.com/lileipisces/NLG4RS

NLG4RS Public
A small ecosystem for Recommender Systems-based Natural Language Generation
☆ 39 · Forks 2

Representative Models

NRT Public
(PyTorch re-implementation) Neural Rating Regression with Abstractive Tips Generation for Recommendation, SIGIR'17
Python · ☆ 10

Att2Seq Public
(PyTorch re-implementation) Learning to Generate Product Reviews from Attributes, EACL'17
Python · ☆ 8

PETER Public
ACL'21 Oral, Personalized Transformer for Explainable Recommendation
Python · ☆ 89 · Forks 20

PEPLER Public
TOIS'23, Personalized Prompt Learning for Explainable Recommendation
Python · ☆ 124 · Forks 20

Datasets

Sentires-Guide Public
A quick guide to Sentires: Phrase-level Sentiment Analysis toolkit, SIGIR'14
Python · ☆ 34 · Forks 5

Tripadvisor

amazon

yelp

Evaluation Metrics

- Text quality
  - BLEU
  - ROUGE
- Explainability
  - Unique Sentence Ratio (USR)
  - Feature Matching Ratio (FMR)
  - Feature Coverage Ratio (FCR)
  - Feature Diversity (DIV)

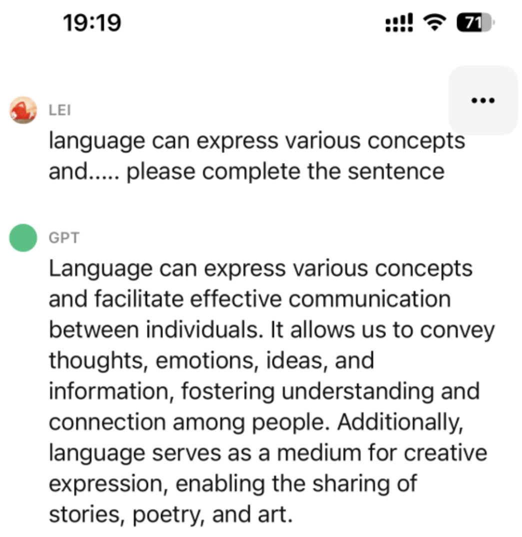
Future Works

LLM-based Recommender Systems

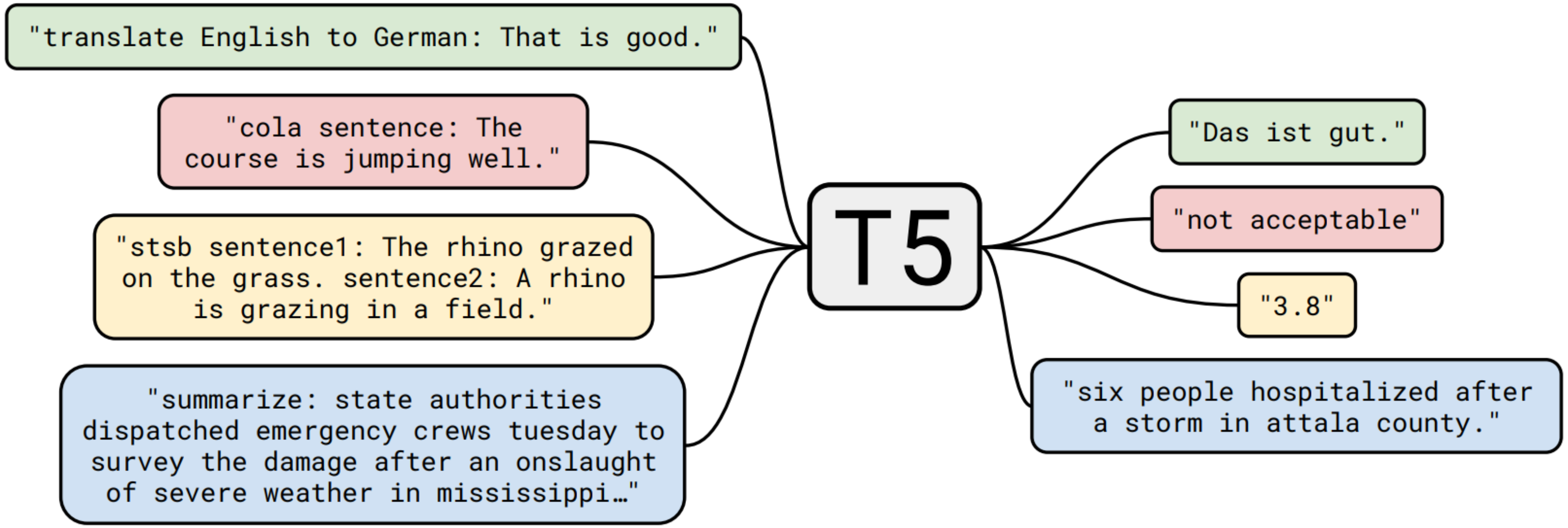
Generative Recommendation

Large Language Models (LLM)

- All tasks formulated as a seq2seq problem
  - Use previous tokens to predict the next token



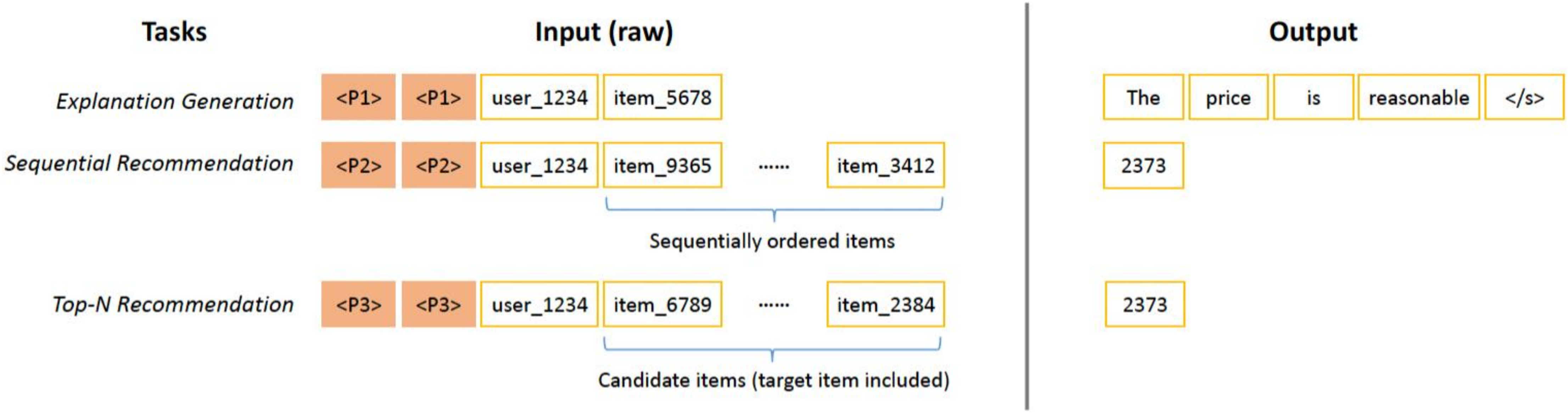
ChatGPT [1]



[1] https://openai.com/chatgpt
[2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR'20.

LLM-based Recommender Systems

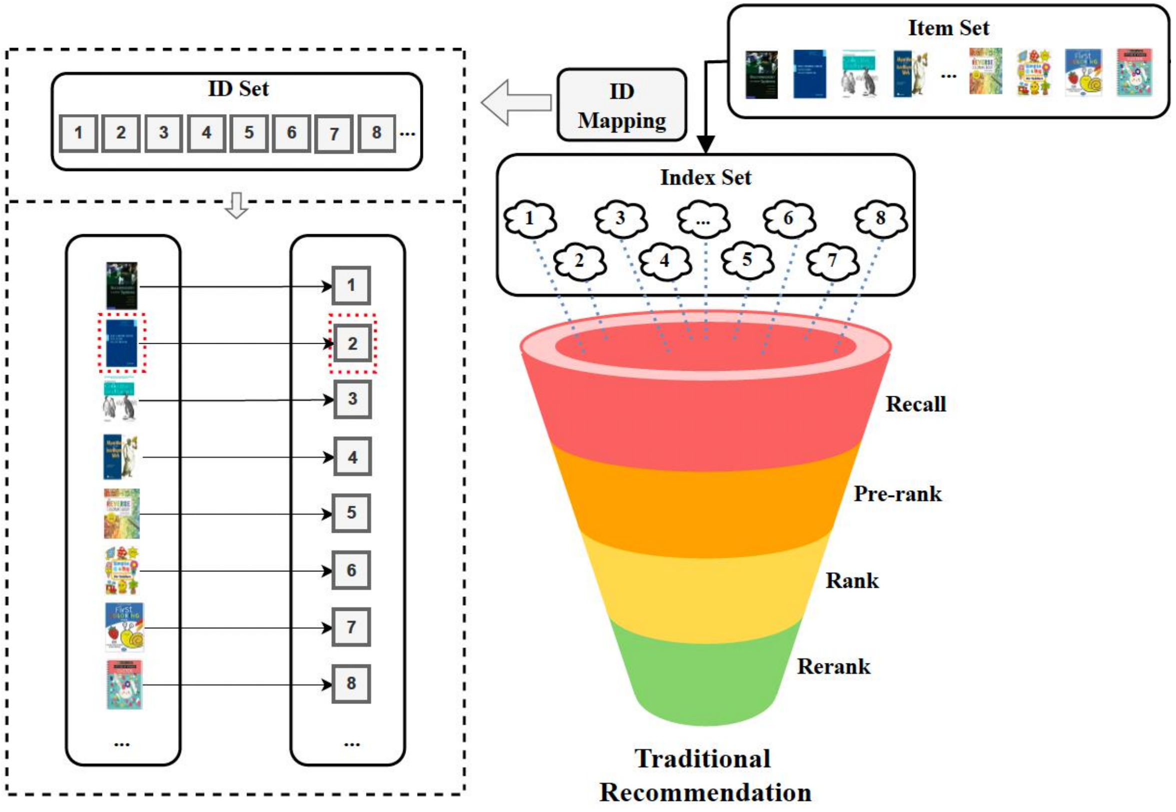
- Recommendation tasks represented as a seq2seq problem
- Multiple tasks integrated into one LLM-based recommendation model



[1] Li, Lei, et al. "Prompt distillation for efficient llm-based recommendation." CIKM'23.

Discriminative Recommendation

- Huge number of items on recommendation platforms
- Computationally expensive score calculation for each item
- Multi-stage filtering to narrow down candidates
  - Simple methods at early stage
  - Complex models at final stage
- Gap between academic research and industrial applications

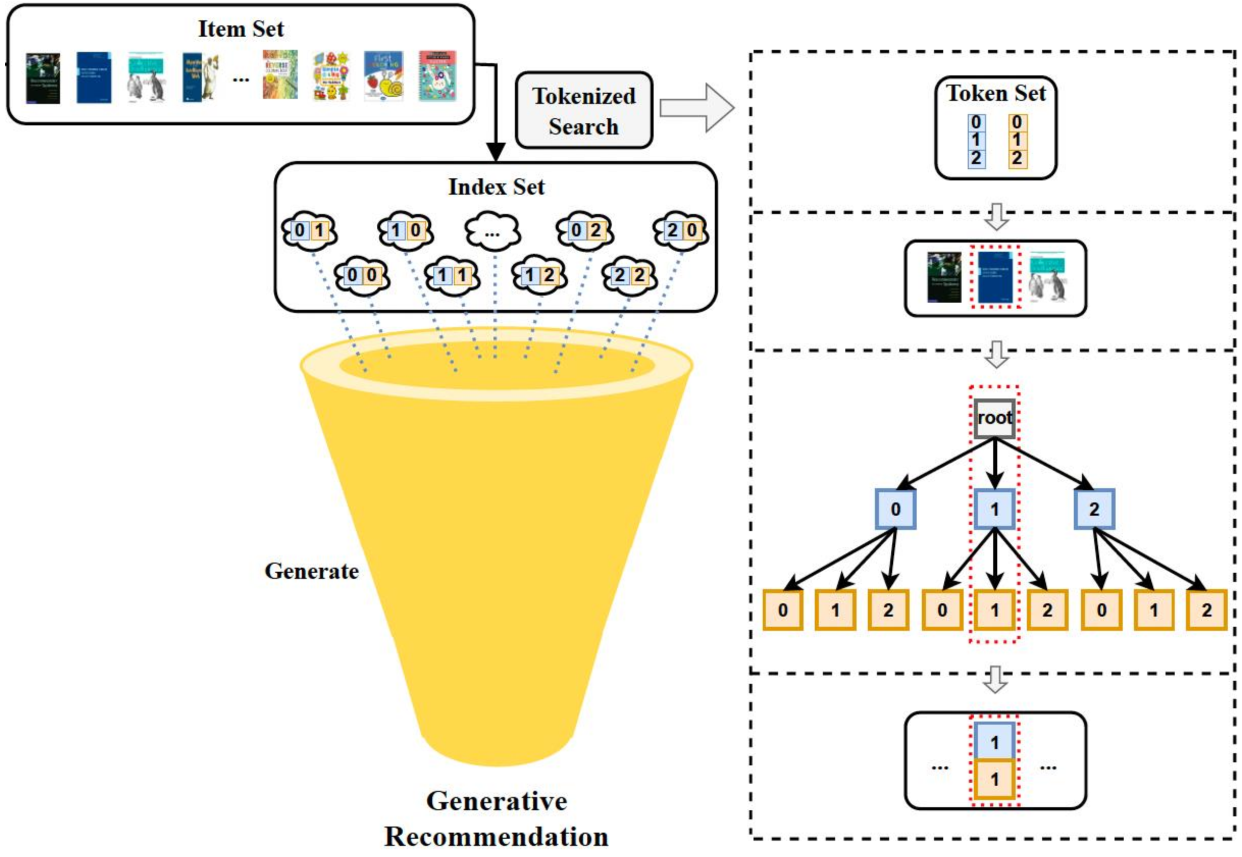


[1] Li, Lei, et al. "Large language models for generative recommendation: A survey and visionary discussions." COLING'24.

Generative Recommendation

- Simplify recommendation process to one stage
  - Directly generate items for recommendation
  - Implicitly enumerate all items

- Use finite tokens to represent infinite items
  - # tokens = 1000
  - ID length = 10 tokens
  - # items = $1000^{10} = 10^{30}$



[1] Li, Lei, et al. "Large language models for generative recommendation: A survey and visionary discussions." COLING'24.

Q&A

Thank you!

csleili@comp.hkbu.edu.hk

lileipisces.github.io