



北京大学

© haoyuan@BDWM

# P大树洞生成器 设计与实现

匡宇轩 陈红韵 王天源

Peking  
University

- 背景

- 项目

- 项目基本框架
- 神经网络结构
- 实验结果及分析
- 生成样例及分析
- 现场测试

- 小结

- 问题和困难

- P大树洞作为校内学生匿名交流平台，是一个完美的语料库（在遵守相关规定的情况下），十分适合NLP相关任务。
- 本项目通过收集大量树洞文本数据，用神经网络进行拟合，希望能创造出一个具有P大学生气质的AI。
- 使用者可以作为“洞主”，模拟发树洞的过程，但是回复的字母君都是AI自动生成。



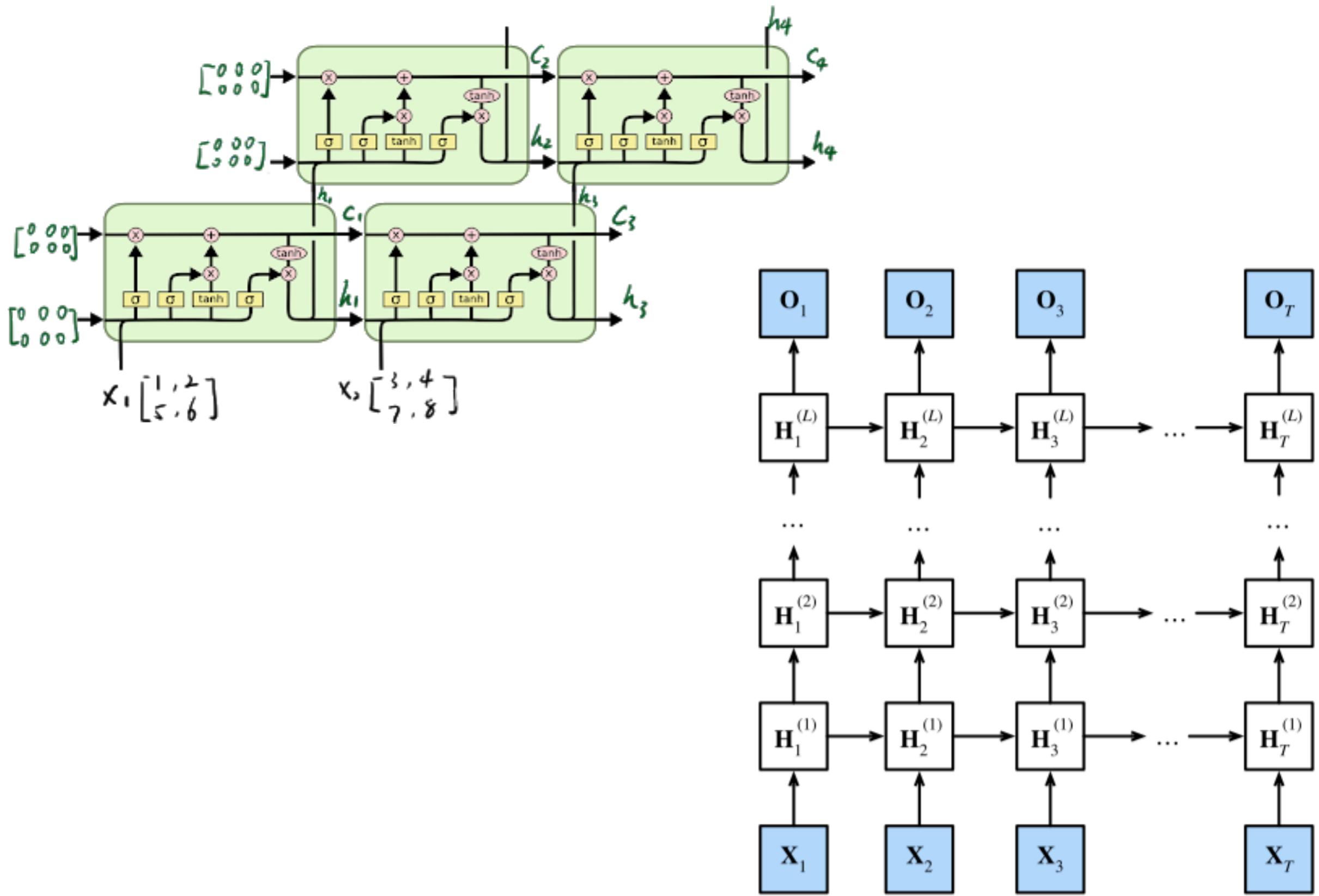


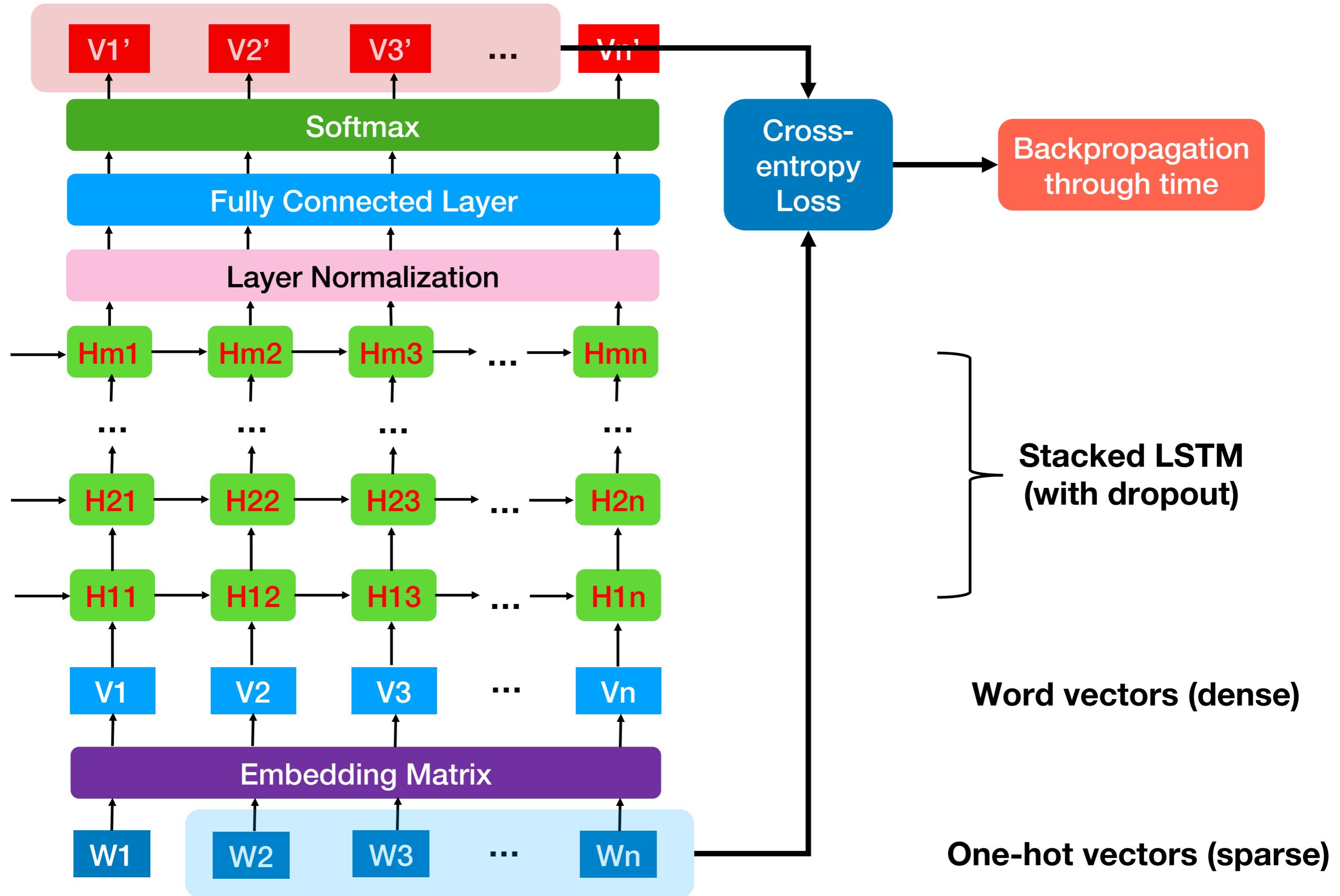
手动收集约**3000**条树洞  
约**一百万字**  
包括日常话题、经典话题、神洞

数据清洗：  
去除洞号、日期、收藏信息  
去除低频字符  
去除过多重复字符，如



## Stacked LSTM





## 50 epochs, 4 hours training on 3080

训练任务

任务模板

任务名称 搜索

搜索

高级搜索

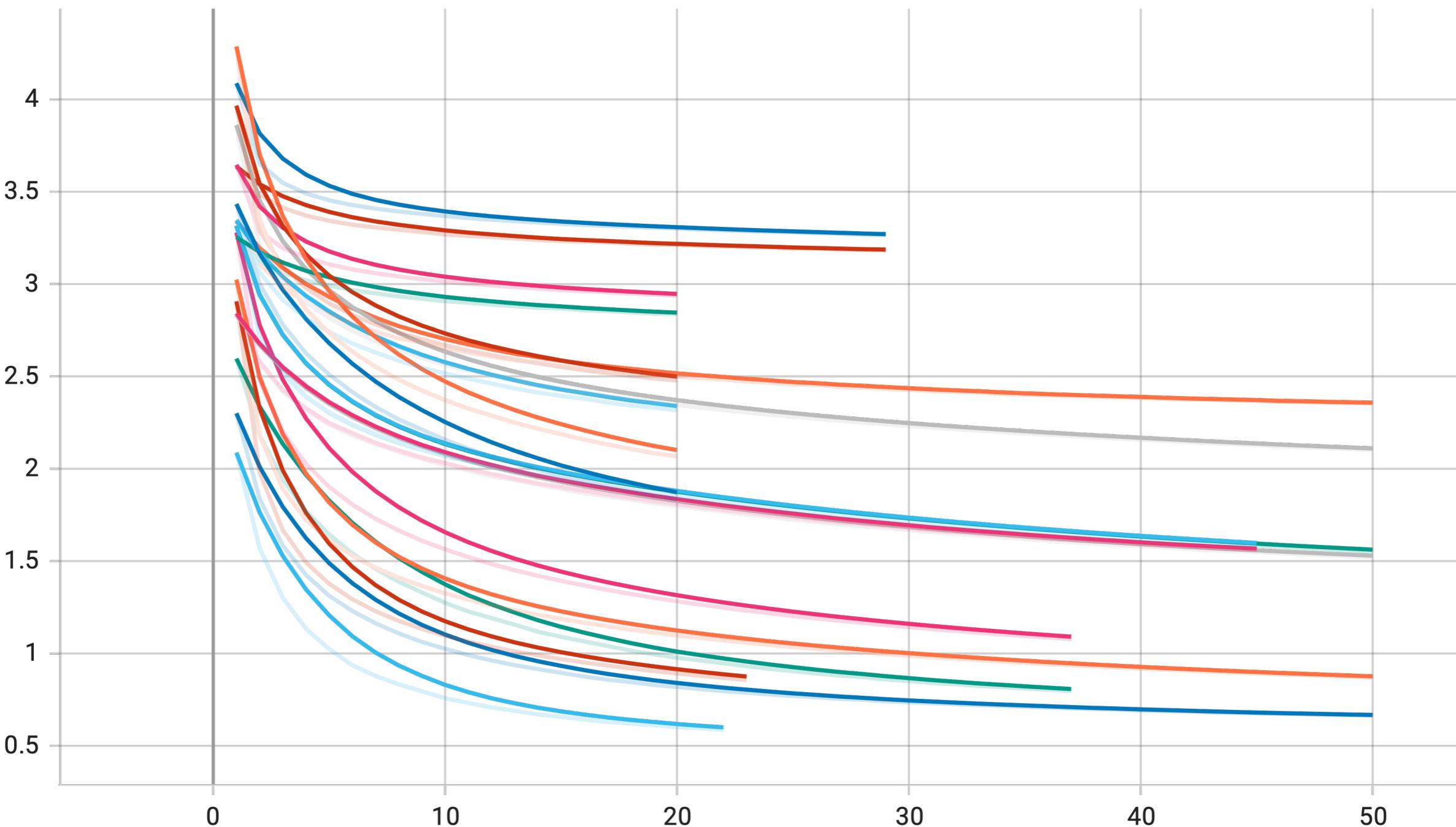
创建任务

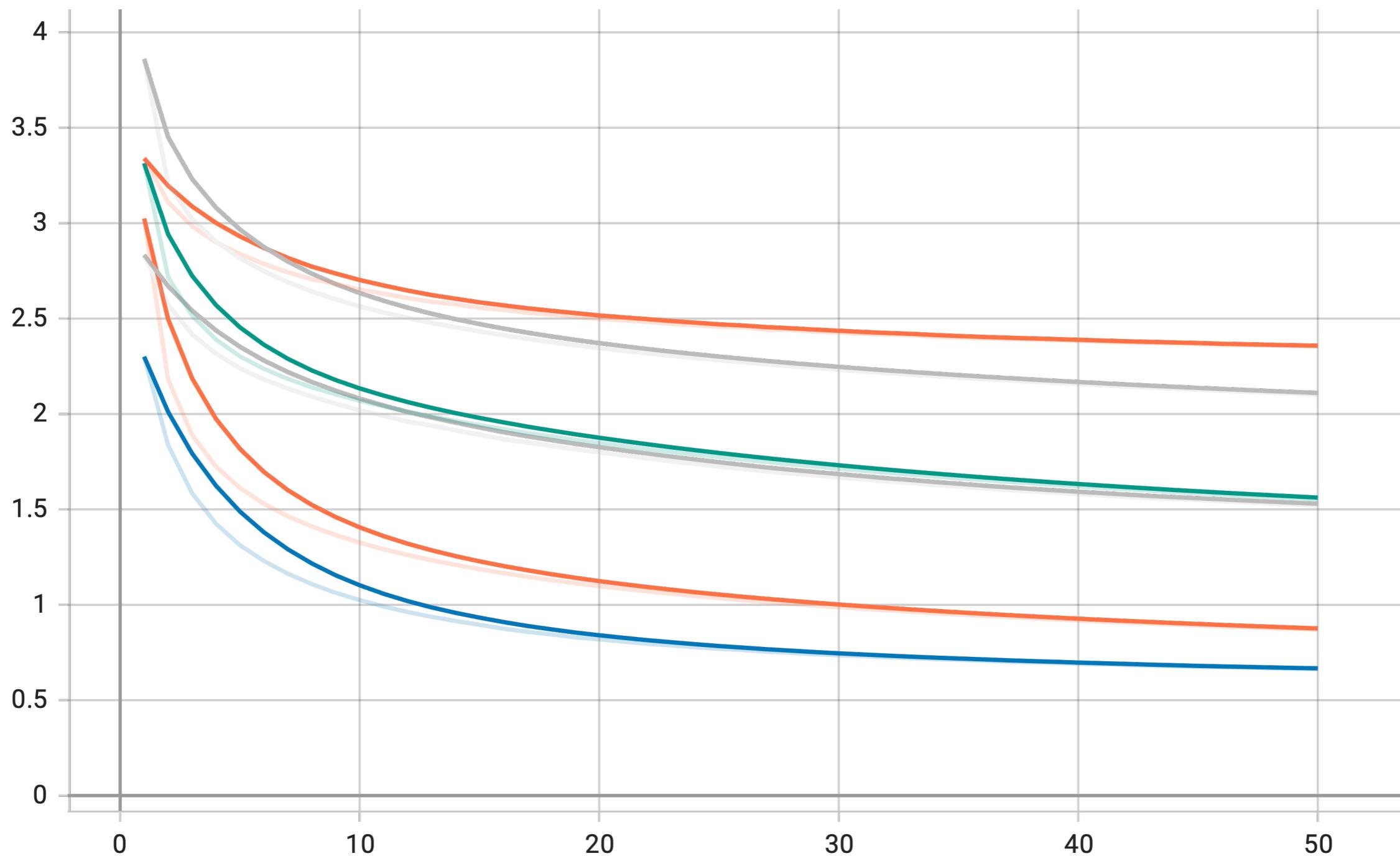
批量删除

<input type="checkbox"/>	任务名称	算法名称	数据集名称	描述	分布式任务	状态	创建时间	运行时长	操作
<input type="checkbox"/>	50_256_3_50	0505	hole_merge		否	● 成功	2022-05-07 22:20:56	3h54m46s	<a href="#">详情</a> <a href="#">删除</a>
<input type="checkbox"/>	50_128_3_50	0505	hole_merge		否	● 成功	2022-05-07 22:15:25	4h9m13s	<a href="#">详情</a> <a href="#">删除</a>
<input type="checkbox"/>	30_512_4_50	0505	hole_merge		否	● 成功	2022-05-07 16:42:33	3h37m27s	<a href="#">详情</a> <a href="#">删除</a>
<input type="checkbox"/>	30_256_4_50	0505	hole_merge		否	● 成功	2022-05-07 16:13:17	3h30m43s	<a href="#">详情</a> <a href="#">删除</a>

Name	Size	Input words	Hidden size	Number of layers	Final val loss
HoleAI-small	4.7MB	50	128	3	1.5476
HoleAI-medium	12.6MB	50	256	3	0.4562
HoleAI-large	37.8MB	30	512	3	0.4354
HoleAI-ultra	46.2MB	30	512	4	0.4640

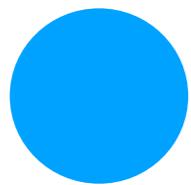
## Overview





**Amazingly, the validation loss is lower than train loss all the time!**

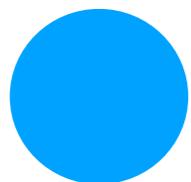
**Losses (smoothed) are monotonically decreasing, which meets our expectation.**



**dropout rate** 不宜过大, **num\_layers** 不宜过多（优化梯度回传）



随机**split**数据集, 可以起到**data augmentation**的作用, 但也因此丢失了部分语义信息



隐层大小至关重要!



[Alice] 表白

[Bob] Bob

[Bob] 来了！

[Carol] 摸摸，因为要留到一个认真的日子！

[Bob] 谢谢A 爱你😘

popi

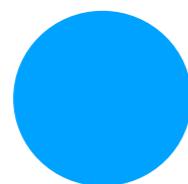
[Alice] 身高体重颜值

[洞主] 170，保密，自我感觉中上（会被人偶尔称赞的程度！）

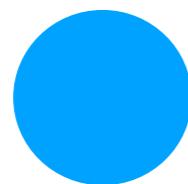
[Bob] Re 洞主: dz是嘉心糖吗

[洞主] Re Bob: 确实

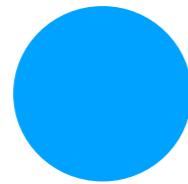
[Alice] Re 洞主: 或者想聊聊也行，我还挺会聊天的



总体来说，捕捉到了语言模式，也有一定的泛化能力，“有内味了”



语义丢失的现象依然存在：  
“自说自话” “指鹿为马” “虚空索敌”



在大段文字的生成上，会出现过拟合现象，  
“金融大师”现身说法

The screenshot shows a web browser window with a dark teal gradient header. The title 'PKU Hole Generator' is centered in large white font. Below it, the text 'Powered by @HoleAI' is displayed. A 'View on GitHub' button is located in a rounded rectangular box. The main content area has a white background. The title 'PKU Hole Generator' is repeated in large green font. Below it are two language selection buttons: 'English' and '中文'. The 'Introduction' section is titled 'Introduction' in green font. It contains text about the project being a course project for 'Introduction to AI(人工智能引论)' at 'Peking University'. It also mentions the project's inspiration from the 'hole(树洞)' at Peking University and its use as a perfect corpus for NLP tasks. The text is in English.

# PKU Hole Generator

Powered by @HoleAI

[View on GitHub](#)

## PKU Hole Generator

English 中文

### Introduction

This is the course project for the NLP class of **Introduction to AI(人工智能引论)** in **Peking University**.

The project was mainly inspired by the [hole\(树洞\)](#) in Peking University, which is a perfect corpus for NLP tasks.

It is designed according to the principle of causal language modeling and implemented in [Python](#) and [PyTorch](#).

The screenshot shows a web browser window with the URL `kryptonite.work` in the address bar. The content area displays two examples of AI-generated dialogue.

**Examples**

Example 1:

😊  
[Alice] 表白  
[Bob] Bob  
[Bob] 来了!  
[Carol] 摸摸，因为要留到一个认真的日子！  
[Bob] 谢谢A 爱你😘

Example 2:

popi  
[Alice] 身高体重颜值  
[洞主] 170，保密，自我感觉中上（会被人偶尔称赞的程度！）  
[Bob] Re 洞主：dz是嘉心糖吗  
[洞主] Re Bob: 确实  
[Alice] Re 洞主：或者想聊聊也行，我还挺会聊天的

**Let's Try!**

🍉

**Model Details @HoleAI**

The screenshot shows a Streamlit web application interface. On the left, there is a sidebar with three sections: "选择一个模型" (Select a model) containing radio buttons for "HoleAI-small", "HoleAI-medium", "HoleAI-large" (which is selected), and "HoleAI-ultra"; "选择一个情绪" (Select an emotion) containing radio buttons for "neutral" (which is selected), "positive", and "negative"; and "选择树洞长度" (Select post length) with a numeric input field set to "5" and minus/plus buttons. The main content area features a large title "P大树洞-爱の引论" (P Big Hole - Love's Introduction) and a subtitle "欢迎来到P大树洞! @HoleAI". Below this is a text input field with placeholder "发一条树洞吧!" (Post a hole here!) and a "开始生成" (Start Generating) button.

选择一个模型

HoleAI-small  
 HoleAI-medium  
 HoleAI-large  
 HoleAI-ultra

选择一个情绪

neutral  
 positive  
 negative

选择树洞长度

5 - +

P大树洞-爱の引论

欢迎来到P大树洞! @HoleAI

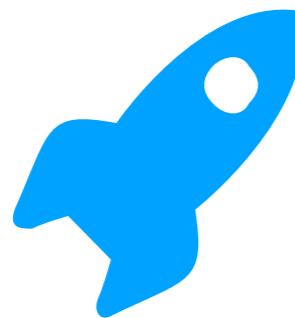
发一条树洞吧!

开始生成

Made with Streamlit

<https://kryptonite.work/pkuhole>

# 小结——意义、问题与困难



- 这个任务十分新颖有趣，而且有一定社会科学方面的研究价值。



- 然而由于数据集的缺乏（根本没有）和数据集不够clean，给我们带来了很大挑战。



- 模型大小与数据集大小的匹配也是一个挑战。



- 不断学习新的技术，看API文档成为家常便饭。

**Thanks for your  
Attention**