北京大学

Peking University

P大树洞生成器设计与实现

匡宇轩 陈红韵 王天源

内容提要

- 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

Background

- 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。

项目——Framework

(1)
- 数据集收集
- 数据预处理

(2)
- 神经网络构建
- 神经网络训练

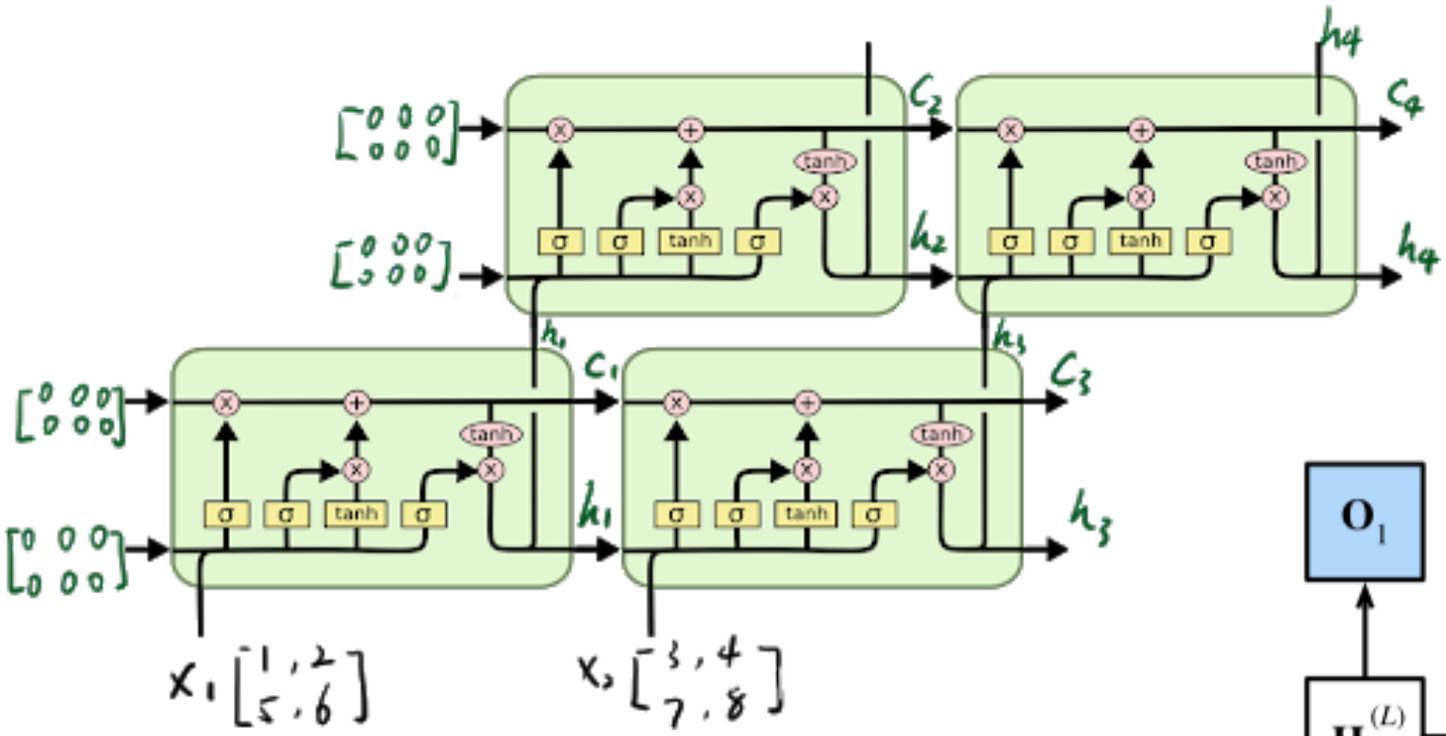
(3)
- 玄学调参 检验结果
- 前端可视化 人机交互

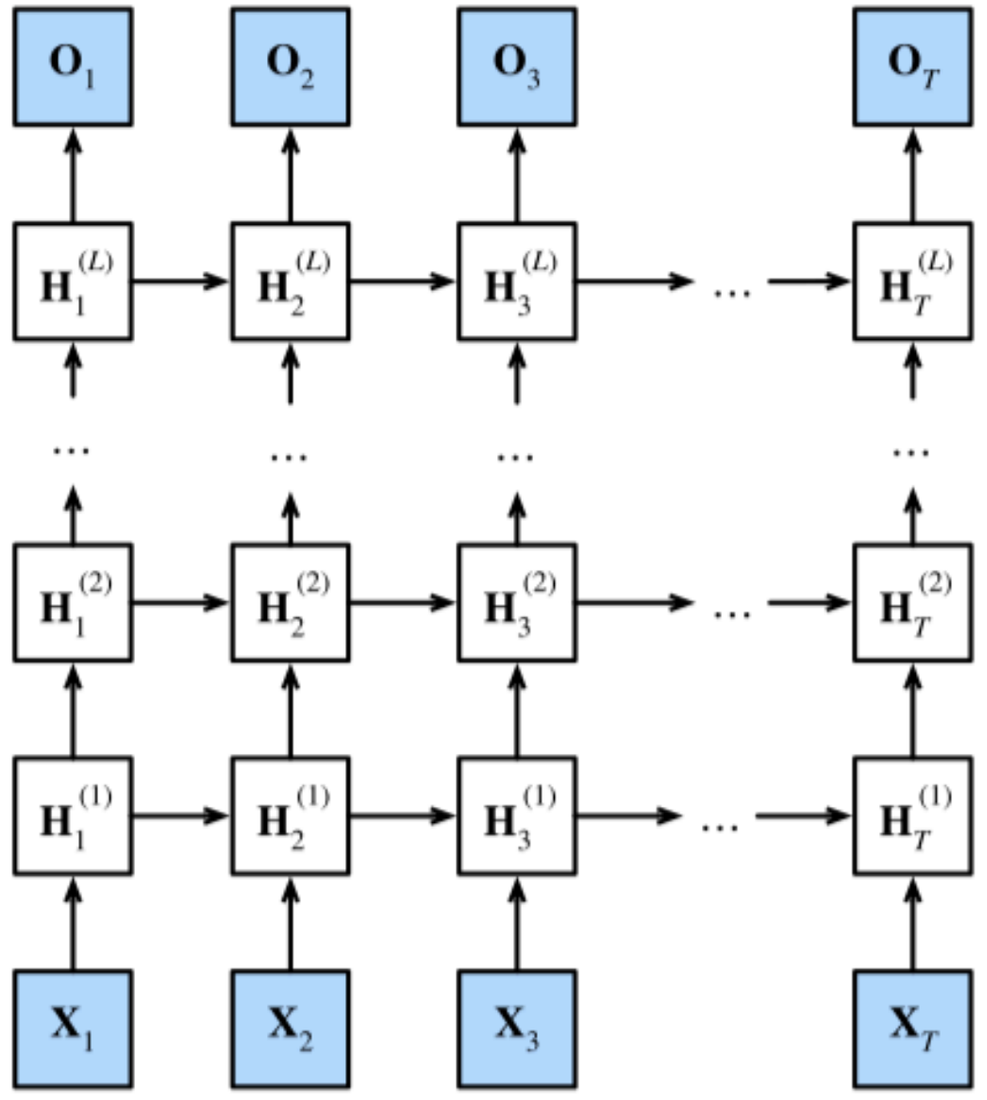
项目——Dataset

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

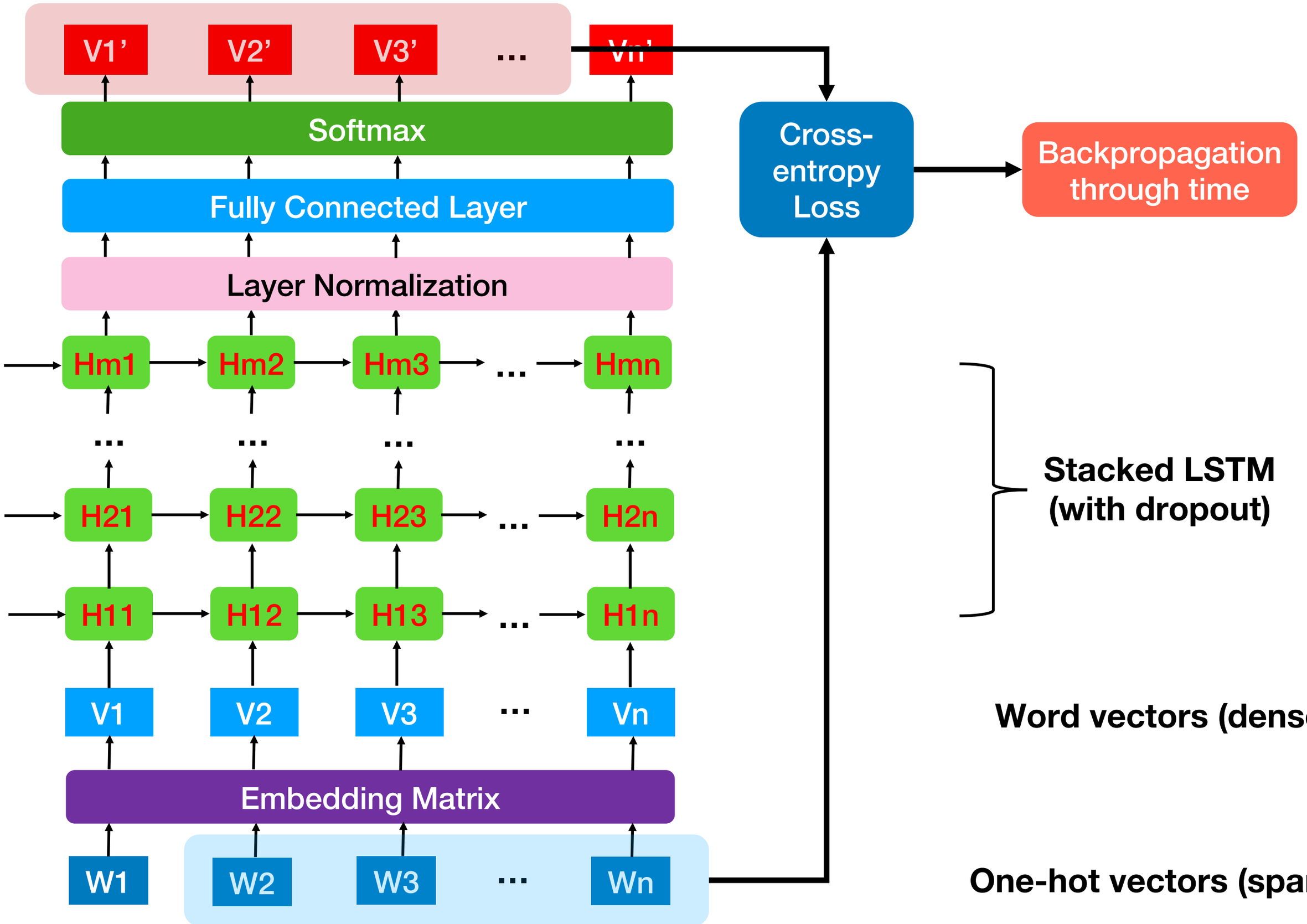
数据清洗：
去除洞号、日期、收藏信息
去除低频字符
去除过多重复字符，如😅

Stacked LSTM





项目——Whole Pipeline



项目——Training

50 epochs, 4 hours training on 3080

训练任务　任务模板

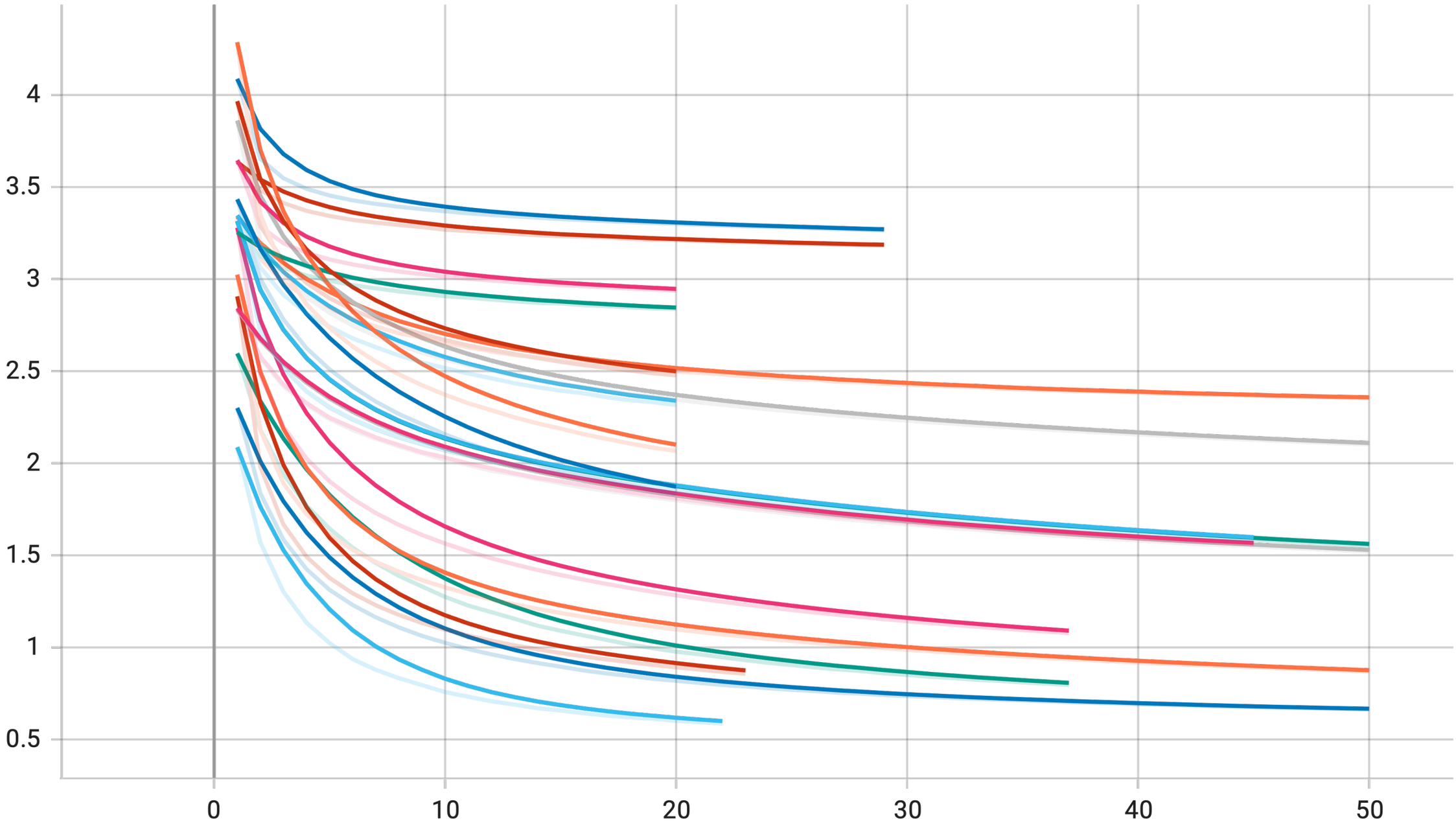
任务名称 搜索　搜索　高级搜索　创建任务　批量删除

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
50_256_3_50	0505	hole_merge		否	● 成功	2022-05-07 22:20:56	3h54m46s	详情 删除
50_128_3_50	0505	hole_merge		否	● 成功	2022-05-07 22:15:25	4h9m13s	详情 删除
30_512_4_50	0505	hole_merge		否	● 成功	2022-05-07 16:42:33	3h37m27s	详情 删除
30_256_4_50	0505	hole_merge		否	● 成功	2022-05-07 16:13:17	3h30m43s	详情 删除

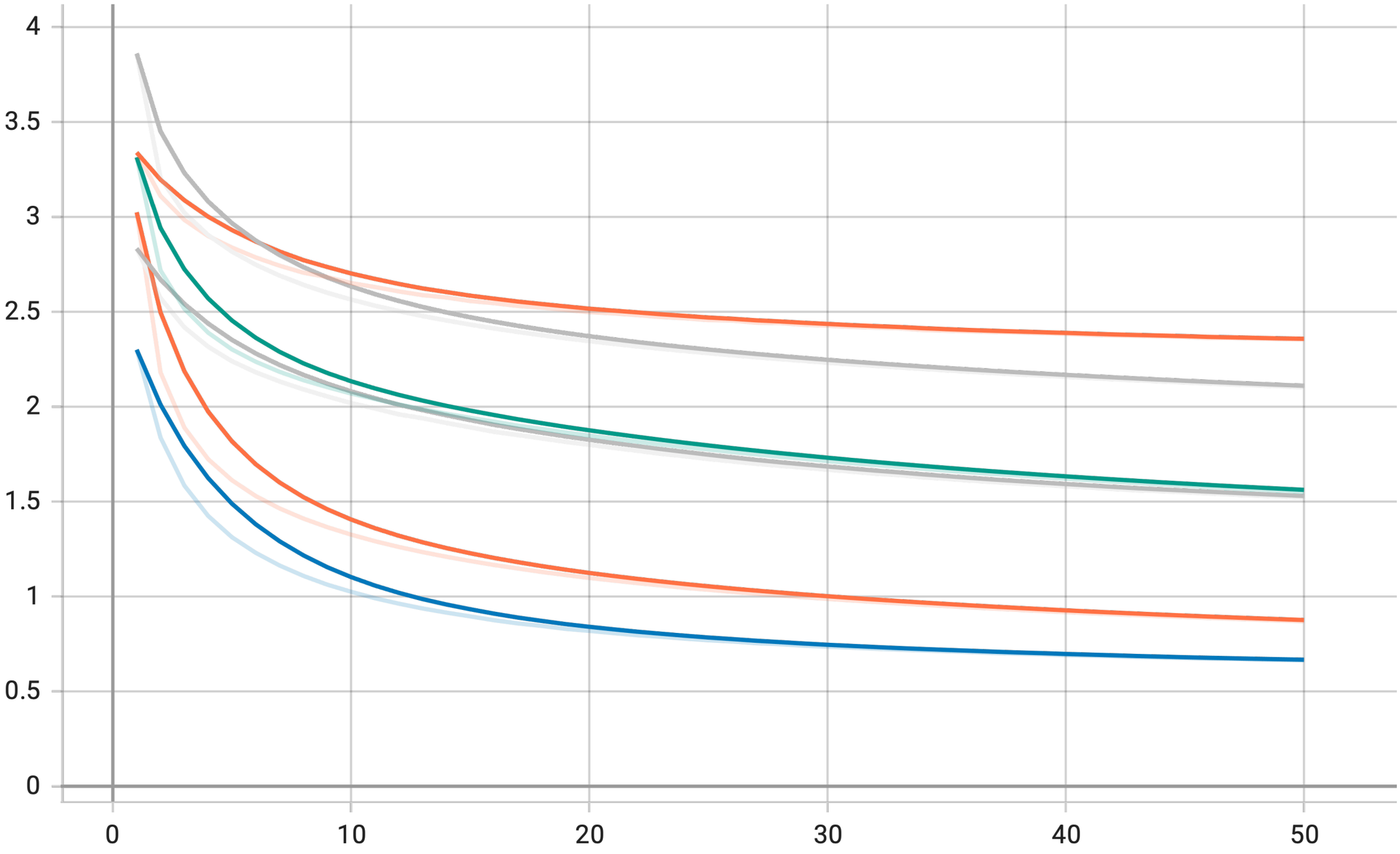
Name	Size	Input words	Hidden size	Number of layers	Final val loss
HoleAI-small	4.7MB	50	128	3	1.5476
HoleAI-medium	12.6MB	50	256	3	0.4562
HoleAI-large	37.8MB	30	512	3	0.4354
HoleAI-ultra	46.2MB	30	512	4	0.4640

项目—— Train & Validation curve

Overview



项目—— Train & Validation curve



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

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

项目——Analysis & Tuning

- **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！

项目——Examples

😅

[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

AIPKU-NLP

项目——Analysis

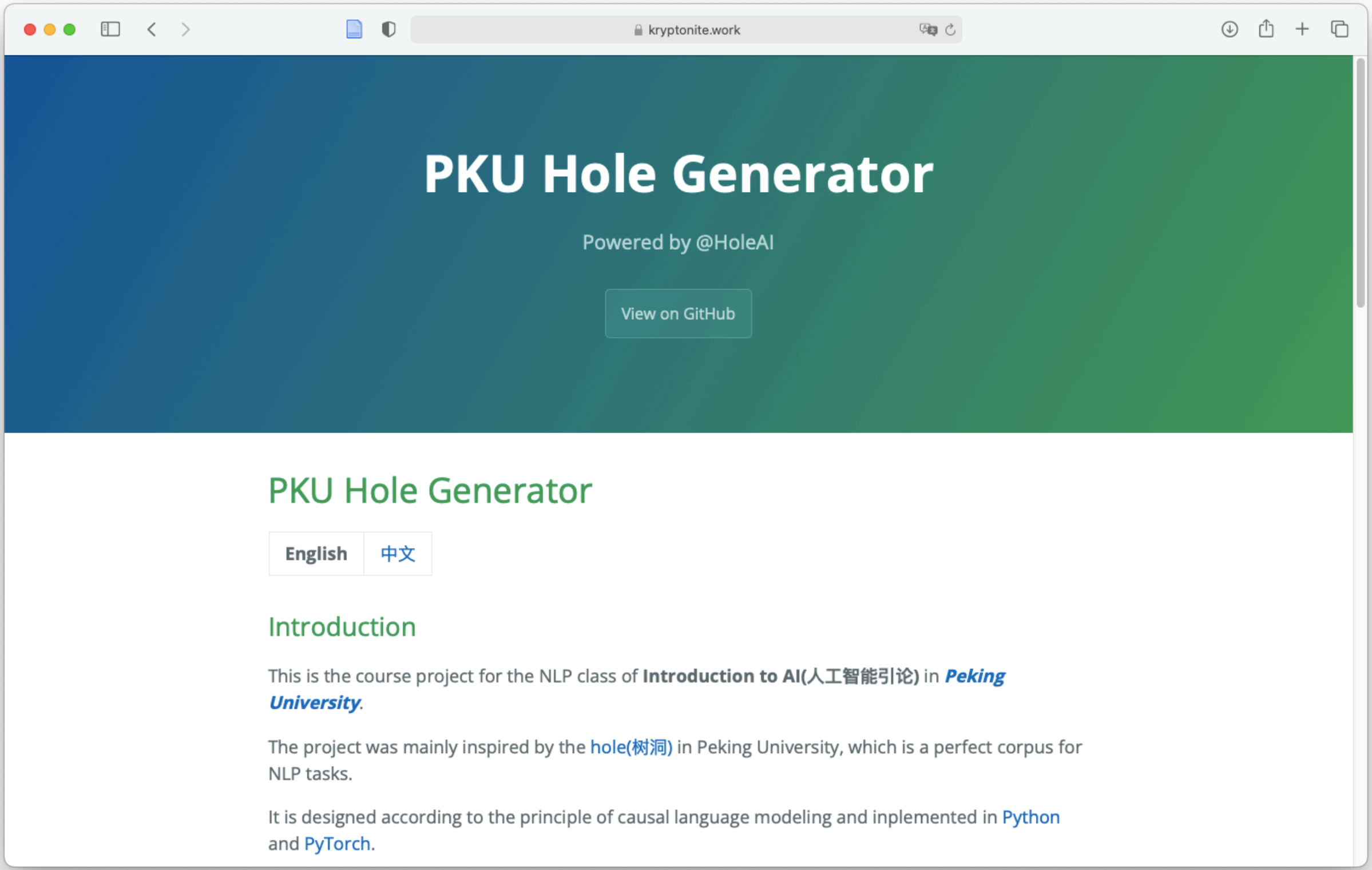
总体来说，捕捉到了语言模式，也有一定的泛化能力，“有内味了”

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

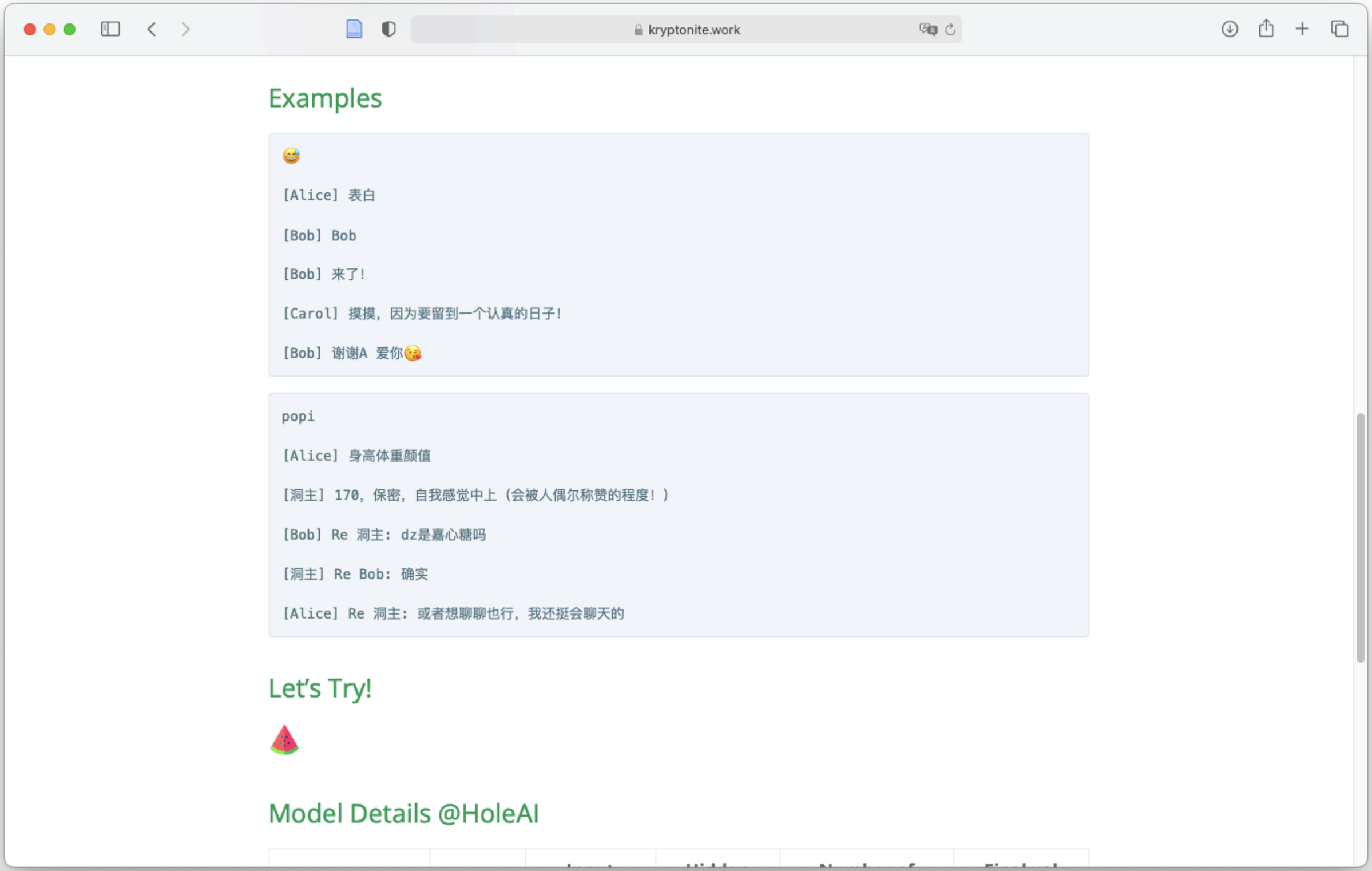
AIPKU-NLP

Web demo & UI design



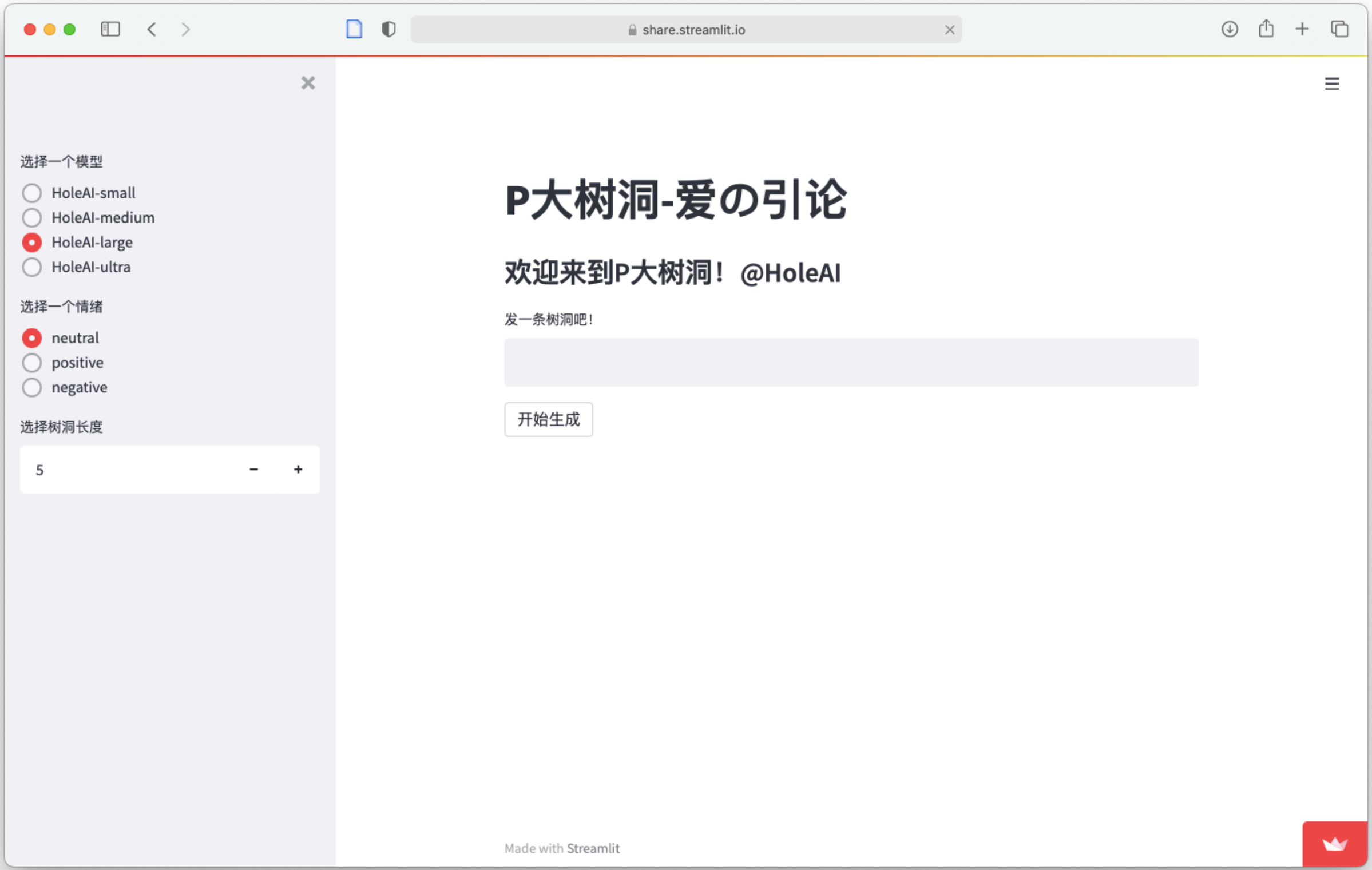
AIPKU-NLP

Web demo & UI design



AIPKU-NLP

Web demo & UI design



AIPKU-NLP

来当洞主吧！（迫真）

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。

Any question?

Thanks for your Attention