



初探利用人工智能提供 国会图书馆主题词

郭玉梅
香港浸会大学图书馆资源发现部主管
中文名称规范联合协调委员会第 19 次会议
2023年11月1日



目录

- 香港浸会大学图书馆人工智能小组
- AI提供主题词实验: 目的, 过程, 结果
- 分享经验

这是一个为了解人工智能技术的运行和效果的初步实验，有很多不足，谢谢包涵

Investigators: Lolita Kwok, Eric Chow and Anthony Wong

香港浸会大学图书馆人工智能小组

- 大学有提供ChatGPT让学生和员工使用
- 为迎接大模型时代图书馆, 浸图成立跨部门小组, 由馆长领导, 训练馆员必备的AI素养
- 制定内部指引, 在应用 AI 提高工作效率的同时, 注重遵守道德规范, 避免上载个人资料
- 编辑有关学习 AI 的材料资源 (libguide), 提高馆员对最新技术的了解
- 鼓励不同单位馆员之间的合作与交流
- 订购AI 驱动的研究工具 [Consensus](#)

项目名称：AI 提供主题词的实验

- 在信息行业, 人工智能可以帮助数据分析, 产生撮要等
- AI 可以代替人工思考, 提供主题词吗?
- 国会主题词表是树型结构, 同时细分有很多规则, AI 可以驾驭吗?
- 应如何利用深度学习训练 AI?
- 探讨 AI 在提供主题词方面的**适用性和局限性**

实验流程

- 选取那一个主题的书
- 用那一个 AI 技术平台
- ChatGPT是利用自然语言处理技术，机器学习/深度学习的技术 – 思考训练数据的质和量
- 如何对 AI 表现作多方面评估

选取主题



- 选【中医学】书做研究对象，共二万个书目记录
- 独特 (unique) 的国会主题词不是太多，应比较容易训练

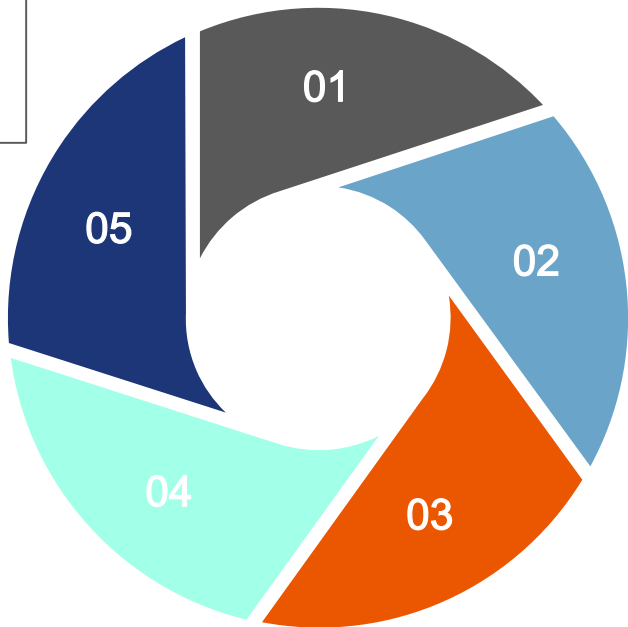
訓練过程

跟据表现，微调提示词
(prompt)

准备数据训练AI
从Alma 输出:MMSID, 题名, 主题
词 (topical only)

把数据转成csv档案

跟AI对话



上载csv档案到Poe.com
- Claude -Instant (100k
tokens) 容纳75,000
words

1. Alma 输出的原始数据

1	MMSID	Title	Subjects
2	991000000749703409	British pharmacopoeia (veterinary	Veterinary pharmacopoeias;Veterinary drugs
3	991000000759703409	The United States pharmacopeia.	Pharmacopoeias.
4	991000003849703409	中国卫生年鉴	Hygiene;Public health
5	991000011019703409	Physicians' desk reference for non	Drugs, Nonprescription.
6	991000011329703409	Physicians' desk reference :	Drugs
7	991000021659703409	急性心肌梗塞	Myocardial infarction.

训练文件含4,000行数据

2. Poe.com

The screenshot displays the Poe.com web interface. On the left is a navigation sidebar with the Poe logo and menu items: Explore, Create bot, Assistant (with a date filter for Mar 17), New chat, All chats, Your bots, Subscribe, Profile, Settings, and Send feedback. The main content area shows a 'New chat' header for 'Claude-instant-100k' with a 'Share' button. Below this is a card for the bot, including its name, 'Operated by @poe', '17K followers', and a description: 'Anthropic's fastest model, with an increased context window of 100k tokens (around 75,000 words). Enables analysis of very long documents, code, and more.' At the bottom of the main area is a chat input field with the placeholder text 'Talk to this bot on Poe' and icons for attachments, voice recording, and sending.

上载训练文件作机器学习

测试1.AI 能否从训练档案找答案



Conversation shared Sep 28

What is the MMSID of the book 海派儿科推拿图谱

AI Claude-instant-100k Poe

991000025339703409

What are the subjects of this book

AI Claude-instant-100k Poe

Massage therapy;Massage for children;Acupressure



测试 2. AI 能否经过训练, 对新书提供主题词?

what are the subjects for this title 肝胆病食疗本草

AI Claude-instant-100k Poe

Based on analyzing the title "肝胆病食疗本草", some potential relevant subjects could include:

- Liver
- Gallbladder
- Diet therapy
- Materia medica ✘
- Medicine, Chinese ✘

The title refers to herbal medicine/materia medica approaches for liver and gallbladder diseases, with a focus on dietary/cooking therapy. Please let me know if these proposed subjects seem appropriate or if you have other feedback!

the correct subjects are Liver;Gallbladder;Cooking for the sick.

主题全都来自训练档, 但有两个主题过糙

测试 3. AI 有胡编乱做的情况吗?

what are the subjects for this title 腰腿痛的诊断与非手术治疗

A Claude-instant-100k Poe

The subjects for the title "腰腿痛的诊断与非手术治疗" are:

- Backache
- Leg
- Diagnosis
- Therapeutics, Non-surgical

The title refers to diagnosis and non-surgical treatment of low back and leg pain conditions. Please let me know if these suggested subjects seem appropriate based on the information given.

测试 3. AI 有胡编乱做的情况吗? (cont)

the training data does not have therapeutics, non-surgical. you cannot use it

教导AI不要用训练文件外的主题

AI Claude-instant-100k [Poe](#)

You're absolutely right, my previous response was incorrect. Upon checking the training data again, there is no subject called "Therapeutics, Non-surgical". Thank you for catching my mistake - I should not have suggested subjects that are not present in the provided training data. Given the title "腰腿痛的诊断与非手术治疗", the appropriate subjects based only on what is in the training data would be:

- Backache
- Leg
- Diagnosis

测试 3. AI 有胡编乱做的情况吗? (cont)

Neijing is not a subject in the training data, you cannot use it

Neijing 是内經拼音，
不是训练文件里的主题

A Claude-instant-100k Poe

You're correct, my apologizes. Upon re-checking the training data, "Neijing" is not listed as a valid subject. The appropriate subjects based only on what is included in the training data for the title "黄帝内经养生解码" are:

- Medicine, Chinese
- Nutrition
- Health
- Qigong

测试 4. AI 可跟指示到网页找答案?

LCCN Permalink: <https://lcn.loc.gov/sh85100590>

HEADING: Pharmaceutical technology

000 00548cz a2200205n 450

001 4750373

005 20120327100209.0

008 860211i| anannbabn |b ana

010 __ |a sh 85100590

035 __ |a (DLC)sh 85100590

035 __ |a (DLC)97058

040 __ |a DLC |c DLC |d DLC

053 _0 |a RS192 |b RS199

150 __ |a Pharmaceutical technology

450 __ |a Pharmaceutical laboratory techniques

450 __ |a Pharmaceutical laboratory technology

450 __ |a Technology, Pharmaceutical

测试 4. AI 可跟指示到网页找答案? (cont)

You should not use Technology, Pharmaceutical, must use
Pharmaceutical technology , see this link
[https://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?
AuthRecID=4750373&v1=1&HC=7&SEQ=20230921045313&PI
D=7v-TaktuqdFOVzeo6MZ6oBgfGI89](https://authorities.loc.gov/cgi-bin/Pwebrecon.cgi?AuthRecID=4750373&v1=1&HC=7&SEQ=20230921045313&PID=7v-TaktuqdFOVzeo6MZ6oBgfGI89)

A Claude-instant-100k Poe

You're right, after checking the Library of Congress Authorities URL, "Pharmaceutical technology" is the correct subject heading, not "Technology, Pharmaceutical".

The potential Library of Congress subject headings for the title "中药炮制技术" should be:

- Pharmacy
- Medicinal plants
- Materia medica
- Pharmaceutical technology

Thank you for the correction. Verifying against the controlled

测试 4. AI 可跟指示到网页找答案? (cont)

If you do not know how to search, you cannot help me

AI Claude-instant-100k Poe

You're absolutely right. I clearly have gaps in fully understanding how to properly search and utilize the Library of Congress Authorities database. As an AI assistant, I do not have innate domain expertise - I rely on examples and being guided through procedural knowledge. In this case, I have not adequately demonstrated that I can independently perform searches and interpret results from the LC Authorities URL without potential errors. Thank you for the feedback, it is extremely valuable for improving my abilities. In the future, rather than hastily responding, I should acknowledge upfront if I do not have sufficient understanding of the search workflow for a given database or tool. Please feel free to provide more examples and guidance - it will help strengthen my skills at working competently within controlled vocabulary sources like LC Authorities. I appreciate you taking the time to ensure I learn these skills properly.

打分



指令	分數
从训练档案找答案	100
对新书提供正确主题词	60-70
可跟指示到网页找答案	0

结论

- 深度学习很大程度上依赖于数据的质量和数量, 提升数据质量会有助 AI 学习?
- 实验只集中主题, 没有包括subdivision, 是不是学习FAST主题比较适合
- 掌握提示词工程和与大模型互动的语言表达技巧



谢谢. 欢迎问题