Nutrition LLM Project

Progress Report (June 10)

Hon Kwan Shun Quinson

Sections

- Data Collection Methods
  - X/Twitter
  - Xiaohongshu
  - Zhihu
- Literature Review
  - OpenPath
  - SkinGPT
  - ChatDiet

Data Collection

Data Collection Summary

- Following initial literature review, social media seems to be a useful source for collecting image-text pairs
- X/Twitter has been used in related research, while Chinese sites including Xiaohongshu and Zhihu have potential
- Data can be accessed either through API or web scraping

X/Twitter

- Need token to get API access and post data
- API access to tweets (GET_2_tweet) is paid
  - Free: No general tweet access with API (can only retrieve current user's posts)
  - 100 USD/month: 10,000 posts/month (may not be sufficient for project)
  - 5000 USD/month: 1,000,000 posts/month (may be too costly)
- Web scraping not allowed in ToS
  - You may not access the Services in any way other than through the currently available, published interfaces that we provide. For example, this means that **you cannot scrape the Services**, try to work around any technical limitations we impose, or otherwise attempt to disrupt the operation of the Services.
- Scraper: https://github.com/kaixinol/twitter_user_tweet_crawler (not tested yet)

Ref: https://developer.x.com/en/docs/twitter-api/rate-limits, https://x.com/en/tos

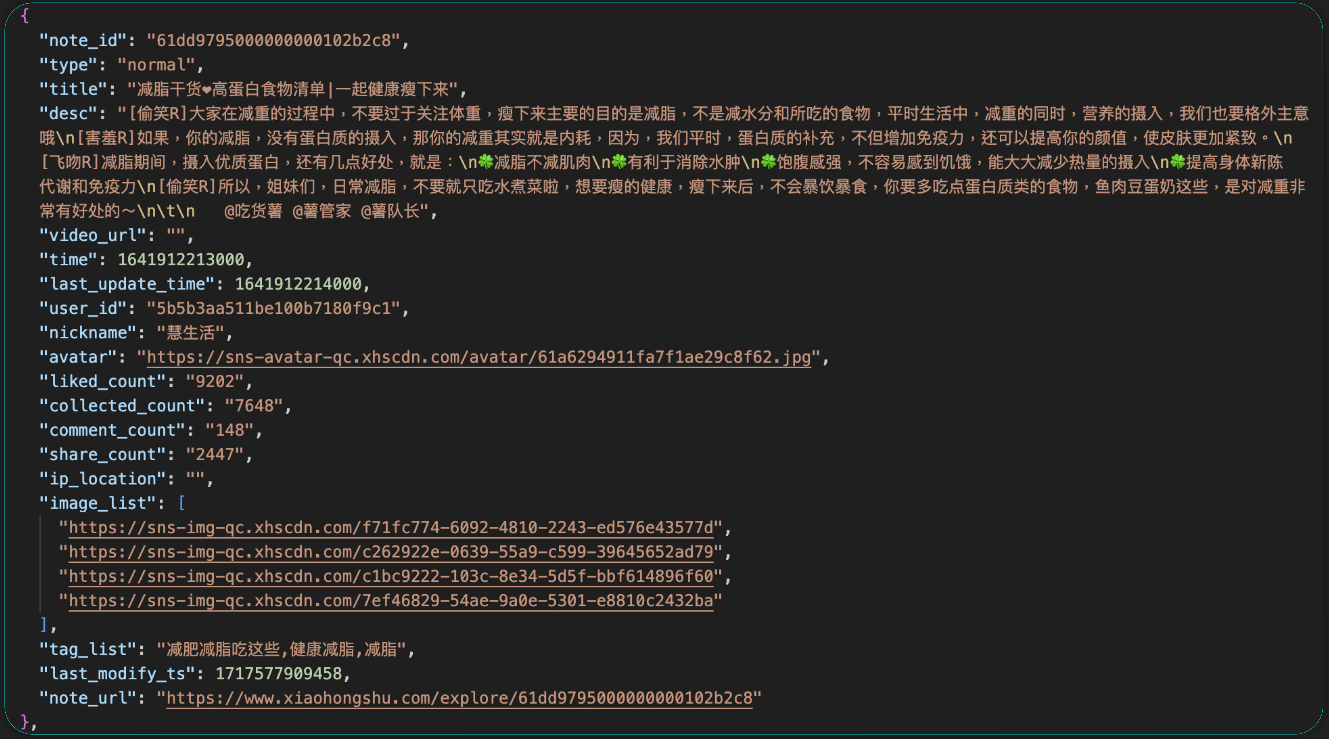
Xiaohongshu

- API exists, but there are problems
  - not sure if it can get notes data, as API is mainly for business automation
  - Developer account verification requires registration as a third-party developer (requiring company details), or as an enterprise on 千帆 which incurs cost
- Usage policy disallows recreation of content and unapproved third-party operations
  - 未经小红书书面允许，用戶不得为任何目的擅自使用、复制、再造这些内容、或创造与内容有关的派生产品。
  - 不得从事下列行为：
    - 以任何方式（包括但不限于盗链、冗余盗取、非法抓取、模拟下载、深度链接、假冒注册等）直接或间接**盗取小红书平台的视频、图文、用戶信息等信息内容**；
    - 通过**非小红书公司开发、授权、许可的第三方软件、插件、外挂、系统，登录或使用小红书平台**，或对小红书平台的正常运行进行干扰、破坏、修改或施加其他影响；
    - 其他以任何不合法的方式、为任何不合法的目的、或与小红书公司为此制定的其他规范和标准不一致的方式使用小红书平台。

Ref: https://agree.xiaohongshu.com/h5/terms/ZXXY20220331001/-1

Xiaohongshu Scraping

- Scraper: https://github.com/NanmiCoder/MediaCrawler
- Works fine currently
- Repo is frequently updated
- Many images -> save to cloud/local storage, or download during training?
- Scraping may infringe on usage policy -> may not be suitable for publications?



```
{
  "note_id": "61dd9795000000000102b2c8",
  "type": "normal",
  "title": "减脂干货❤高蛋白食物清单|一起健康瘦下来",
  "desc": "[偷笑R]大家在减重的过程中，不要过于关注体重，瘦下来主要的目的是减脂，不是减水分和所吃的食物，平时生活中，减重的同时，营养的摄入，我们也要格外主意
  哦\n[害羞R]如果，你的减脂，没有蛋白质的摄入，那你的减重其实就是内耗，因为，我们平时，蛋白质的补充，不但增加免疫力，还可以提高你的颜值，使皮肤更加紧致。\n
  [飞吻R]减脂期间，摄入优质蛋白，还有几点好处，就是：\n🍀减脂不减肌肉\n🍀有利于消除水肿\n🍀饱腹感强，不容易感到饥饿，能大大减少热量的摄入\n🍀提高身体新陈
  代谢和免疫力\n[偷笑R]所以，姐妹们，日常减脂，不要就只吃水煮菜啦，想要瘦的健康，瘦下来后，不会暴饮暴食，你要多吃点蛋白质类的食物，鱼肉豆蛋奶这些，是对减重非
  常有好处的~\n\t\n   @吃货薯 @薯管家 @薯队长",
  "video_url": "",
  "time": 1641912213000,
  "last_update_time": 1641912214000,
  "user_id": "5b5b3aa511be100b7180f9c1",
  "nickname": "慧生活",
  "avatar": "https://sns-avatar-qc.xhscdn.com/avatar/61a6294911fa7f1ae29c8f62.jpg",
  "liked_count": "9202",
  "collected_count": "7648",
  "comment_count": "148",
  "share_count": "2447",
  "ip_location": "",
  "image_list": [
    "https://sns-img-qc.xhscdn.com/f71fc774-6092-4810-2243-ed576e43577d",
    "https://sns-img-qc.xhscdn.com/c262922e-0639-55a9-c599-39645652ad79",
    "https://sns-img-qc.xhscdn.com/c1bc9222-103c-8e34-5d5f-bbf614896f60",
    "https://sns-img-qc.xhscdn.com/7ef46829-54ae-9a0e-5301-e8810c2432ba"
  ],
  "tag_list": "减肥减脂吃这些,健康减脂,减脂",
  "last_modify_ts": 1717577909458,
  "note_url": "https://www.xiaohongshu.com/explore/61dd9795000000000102b2c8"
},
```

Zhihu

- Usage policy: 您应对您使用本平台的行为负责，除非法律允许或者经知乎事先书面许可，您使用本平台不得具有下列行为：
  - 利用或针对本平台进行任何危害计算机网络安全的行为，包括但不限于：
    - 以任何方式（包括但不限于盗链、冗余盗取、**爬取、抓取**、模拟下载、深度链接、假冒、模拟注册等）直接或间接**盗取知乎平台的数据和内容**；恶意注册知乎账号，包括但不限于频繁、批量注册账号；
    - 违反法律法规、本协议、本平台的相关规则及侵犯他人合法权益的其他行为。
- Scrapers: https://github.com/search?q=zhihu+pushed%3A%3E2023-01-01&type=Repositories&ref=advsearch&l=&l=
  (many search results seem to be not maintained/not working, will test them later)
- I can try writing a custom Selenium script? (Don’t know how to pass authentication though, have not needed to do that before)

Ref: https://www.zhihu.com/term/zhihu-terms

Other Scrapers

- https://gitee.com/AJay13/ECommerceCrawlers
- https://open.onebound.cn/help/api/
  - Cheap-ish custom API
  - Unsure if it works

Other Nutrition Data Sources

- https://www.kaggle.com/datasets/gokulprasantht/nutrition-dataset
    - List of nutritional values of different ingredients
    - Useful for building pipelines like ChatDiet (can go into detail later)

Other text sources

- CommonCrawl
  - Massive dataset (100TB) for pre-training
  - Filtering on nutrition based facts should be doable but would lead a lot more data cleaning
- FineWeb
  - Includes deduplicated dumps of CommonCrawl since 2013
  - Samples of smaller size exist (10B, 100B and 350B tokens) which can be filtered
- Medical Journals
  - ~280,000 articles searched through the term “nutrition” on PubMed
  - Likely to have higher quality information, but may be too technical

Literature Review

OpenPath

- Data: pathology related images from Twitter
- Filtering techniques: English, no retweet, no sensitive info, most likes, not a question, etc.
- Architecture: CLIP (ViT + Transformer)
- Training Method: Contrastive Image-Text Learning
- Demonstrated applications: zero-shot transfer learning, image retrieval from text/image (evaluated with recall@10, recall@50)

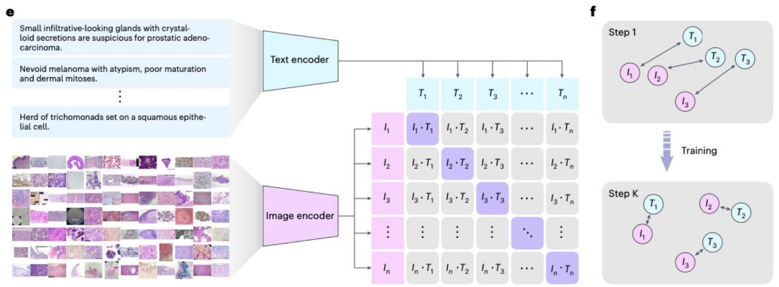
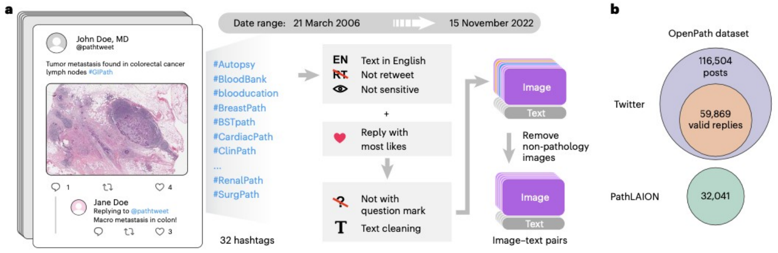


Fig. 1 | Overview of the study. a, Flowchart of data acquisition from medical Twitter. **b**, Overview of the OpenPath dataset. **c**, Total number of available image–text pairs from tweets and replies within each Twitter hashtag (sorted in alphabetical order). Replies are those that received the highest number of likes in Twitter posts, if applicable. **d**, Density plot of the number of words per sentence in the OpenPath dataset. **e**, The process of training the PLIP model with paired image–text dataset via contrastive learning. **f**, Graphical demonstration of the contrastive learning training process.

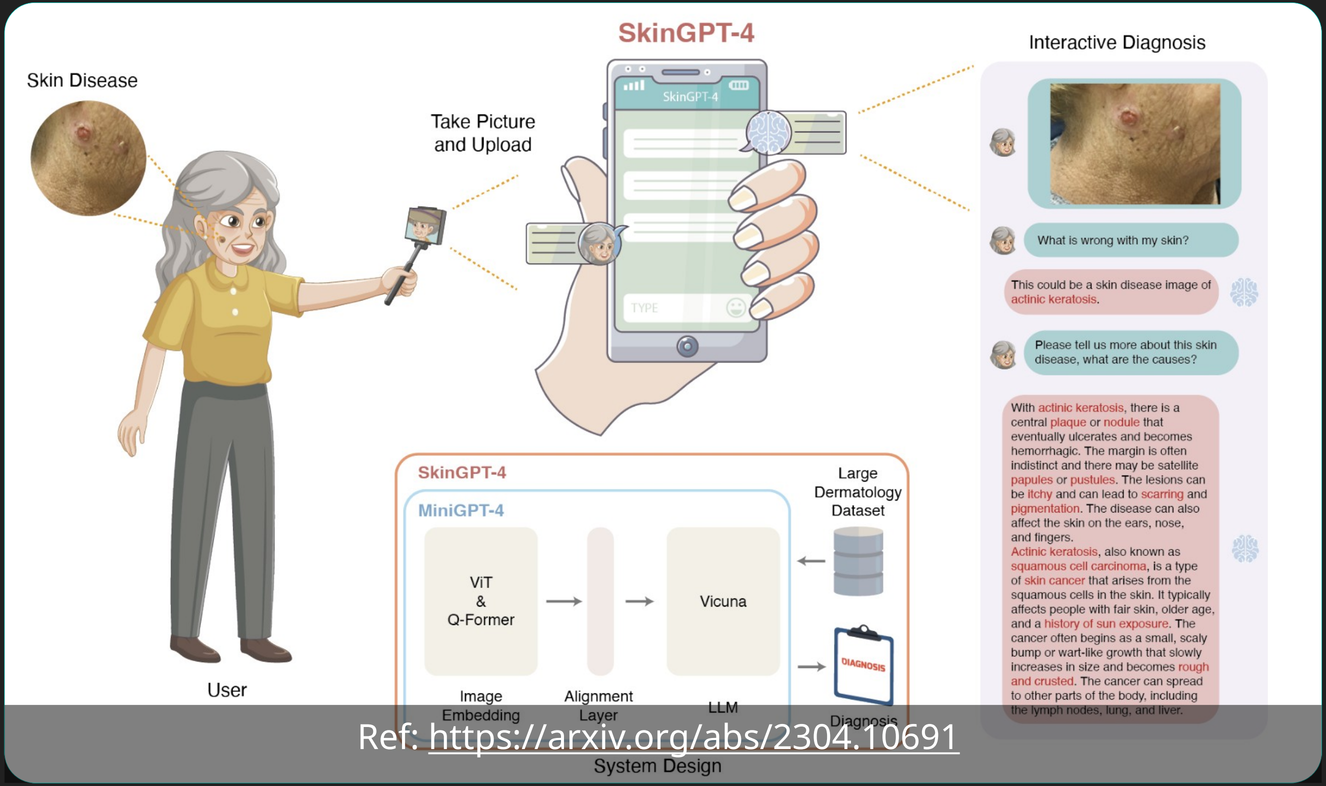
Ref: https://www.biorxiv.org/content/10.1101/2023.03.29.534834v1.full



Ref: https://www.biorxiv.org/content/10.1101/2023.03.29.534834v1.full

SkinGPT (+ MiniGPT)

- Data: public datasets + private dataset (presumably from hospital based on authors)
- Filtering techniques: English, no retweet, no sensitive info, most likes, not a question, etc.
- Architecture: MiniGPT-4 (ViT, Q-former, alignment layer, Vicuna)
- Training Method: LLM Fine-tuning
- Demonstrated applications: Dermatology chatbot (Evaluated quantitatively by field experts)



What is wrong with my skin?

This could be a skin disease image of actinic keratosis.

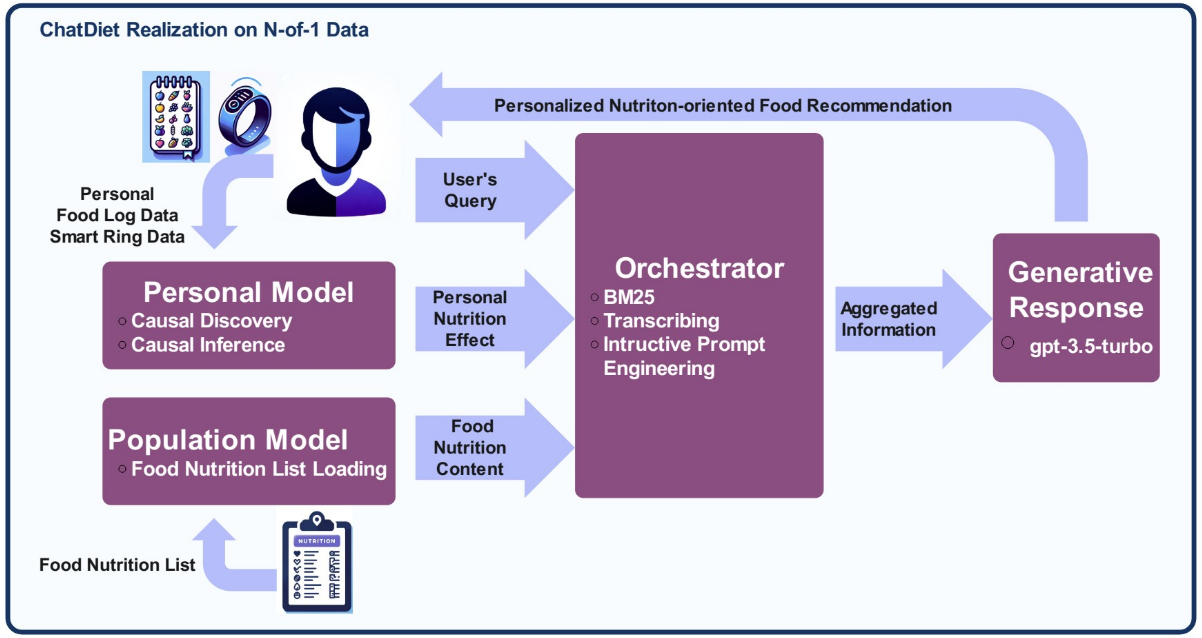
Please tell us more about this skin disease, what are the causes?

With actinic keratosis, there is a central plaque or nodule that eventually ulcerates and becomes hemorrhagic. The margin is often indistinct and there may be satellite papules or pustules. The lesions can be itchy and can lead to scarring and pigmentation. The disease can also affect the skin on the ears, nose, and fingers.
Actinic keratosis, also known as squamous cell carcinoma, is a type of skin cancer that arises from the squamous cells in the skin. It typically affects people with fair skin, older age, and a history of sun exposure. The cancer often begins as a small, scaly bump or wart-like growth that slowly increases in size and becomes rough and crusted. The cancer can spread to other parts of the body, including the lymph nodes, lung, and liver.

Ref: https://arxiv.org/abs/2304.10691

ChatDiet

- Data: Mobile app, wearable device, food nutrition list
- Architecture:
  - Personal model: causal discovery and inference
  - Population model: Databases
  - Orchestrator: BM25 (bag-of-words) retrieval, prompt engineering
  - Generative Model: ChatGPT
- Demonstrated applications: Nutrition chatbot (evaluated manually with accuracy metric)



Ref: https://arxiv.org/abs/234.10691

Other ideas

- Retrieval-augmented generation (RAG)
  - Build vector databases for articles
  - Use article context vectors in LLM generation
  - Simpler pipeline, but requires high-quality chunks to work effectively
- Knowledge graphs
  - Build graphs connecting entities with various relationships
  - Use knowledge graph queries to improve prompt and LLM generation
- Function calling
  - Let the LLM call functions to get relevant information

Takeaways

- Collect data from social media is doable, but it can be costly/inappropriate for publication
- Alternatives may need to considered (e.g. general nutrition/text datasets)
- There are a variety of methods possible for incorporating data into the nutrition LLM
  - Images
  - Databases
  - Causal graphs
  - Other LLM techniques

End of Report