Cancer and diabetes patient phenotyping

-- A large language model (LLM) solution

Hanyin Wang

May 8, 2024

Problem description

- **Large language model Augmented sYmptom Extraction & Recognition (LAYER)**
- **Devise a method to determine whether a note indicates the patient has diabetes and/or cancer.**
  - current or historic
  - “YES” -- positive, “NO” -- negative, or “MAYBE” -- not mentioned
- **Data (N = 2,000 derived from [Asclepius Synthetic Clinical Notes dataset]())**
  - Binary label (0/1 for *both* diabetes & cancer)
  - Textual label (“YES”, “NO”, “MAYBE” for diabetes *or* cancer)
  - No label (over 90%)

Table 1. Case numbers and allocation.

		binary label	textual label <cancer>	textual label <diabetes>	w/o binary label
# patients		50	59	42	1,950
allocation	zero-shot	Evaluation	--		--
	RLHF		Reward model training (step 2)		PPO (step 3)

Study design

1. Zero-shot GPT-4
2. Fine-tuning using Reimforcemnent Learning with Human Feedback (RLHF)
    1) Base (policy) model: mistralai/Mistral-7B-Instruct-v0.2
    2) Reward model fine-tuning
        a. Base model: TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T
        b. Training data: answer pairs derived from the sample dataset.
        c. Resulting reward model: hanyinwang/layer-project-reward-model
    3) Update policy model with PPO
        a. Training data: unlabeled data from provided dataset
        b. Resulting model: hanyinwang/layer-project-diagnostic-mistral

Zero-shot GPT4

- **GPT-4 (gpt-4-0125-preview)**
  - Prompt ➡️
  - Expected output
    {"<condition>": ”YES"}
    {"<condition>": ”NO"}
    {"<condition>": "MAYBE”}

GPT-4 prompt template:
You are a medical doctor specialized in <condition> diagnosis. From the provided document, assert if the patient ***historically and currently*** *has <condition>. For each condition, only pick from "YES", "NO", or "MAYBE". And you must follow the format without anything further. The results have to be directly parseable with python json.loads().*
Sample output: {"<condition>": "MAYBE"}
Never output anything beyond the format.
Provided document: <note>

Table 2. Answer and label correspondence.

answer	label
YES	1
NO	0
MAYBE	

Table 1. Case numbers and allocation.

		binary label	textual label <cancer>	textual label <diabetes>	w/o binary label
# patients		50	59	42	1,950
allocation	zero-shot	Evaluation	--		--
	RLHF		Reward model training (step 2)		PPO (step 3)

- **Performance**
  - All gpt-4 outputs can be found in [this folder]
    - Diabetes 👍
    - Cancer -- 4 misclassified cases 🤨
      - All false negatives

Table 3. Zero-shot GPT-4 performances.

		accuracy	precision	recall
cancer	0	0.92	0.88	1.00
	1		1.00	0.80
diabetes	0	1.00	1.00	1.00
	1		1.00	1.00

Zero-shot GPT4 – Error analysis 🧐

Table 4. Error analysis of all misclassified cases of zero-shot classification using GPT-4.

ID	GPT-4 response	related snippet from note	analysis	conclusion
1814	{"cancer": "MAYBE"}	...She had a past medical history of **chronic lymphocytic leukemia** (in remission). ...	The case was not focused on cancer, but the patient did have a medical history of leukemia.	GPT is **wrong**
3146	{"cancer": "MAYBE"}	...The mass was diagnosed as **LCH**. ...	In the original case report, the following points were made: "Lobular capillary hemangioma is a rare, rapidly growing, **benign tumor**", meaning that LCH is not considered as cancer. However, this information was not included in the note. Furthermore, according to NIH, "It is not known whether LCH is a form of cancer or a cancer-like disease." Therefore, it is a tough case and GPT-4 might make a correct judgement in this case.	GPT is **right**
2117	{"cancer": "NO"}	... Tumor histology was reminiscent of **desmoid fibromatosis** and consistent with **desmoplastic fibroma**. ...	Both are **benign** conditions. The original case report presented two rare bone tumors cases, one was benign (presented in the note), the other was malignant (not presented in the note).	GPT is **right**
2840	{"cancer": "NO"}	... "The primary tumor was diagnosed as a **glomus tumor** based on these findings. " ...	The glomus tumors reported have been mostly benign neoplasms and **very rarely malignant**. The discussion over the malignancy of the tumor was not included in this summary. According to the additional description in the original report: "...This lesion met the malignancy criteria for size, but the tumor had low malignant features of low mitotic activity and absence of significant nuclear atypia. And no recurrence of 8 years was the **basis for the lesion being benign**. ..."	The conclusion of GPT is **correct**, but the provided note lacks necessary details.

Zero-shot GPT4 -- Summary

- **Overall, very strong zero-shot performance 💪**
  - Neat & directly parseable with python
  - Overlooked historic condition
- **Proposed improvements**
  - Prompt engineering: The model sometimes ignores the instruction on "historically and currently", we could ask about the two statuses in individual prompts.
  - Chaining the document: it might also be possible that the model overlooked the provided document, in this case, instead of stuffing, we could use map-rerank to make the model carefully go through each piece of provided information.
    - Current input are short (max token length = 632 for gpt-4), map-rerank could also be considered when dealing with longer inputs
  - Instead of only textual generation, we could also let the model output a probability alongside. This probability can be used for binary classification with a proper threshold, or even for multi-class classification.
  - Ground truth annotation: trichotomized label v.s. dichotomized label.
  - Additional case review: three of the false negative cases (3146, 2117, and 2840) need additional review and possibly annotation update.
  - Domain specific fine-tuning

RLHF Fine-tuning

Table 1. Case numbers and allocation.

		binary label	textual label <cancer>	textual label <diabetes>	w/o binary label
# patients		50	59	42	1,950
allocation	zero-shot	Evaluation	--		--
	RLHF		Reward model training (step 2)		PPO (step 3)

- Base (policy) model: mistralai/Mistral-7B-Instruct-v0.2
  - Supervised fine-tune (STF) – skipped
- Reward model fine-tuning
  - Base model: TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T
  - Training data: answer pairs derived from the given dataset, available at hanyinwang/layer-project-reward-training (n = 101)
    - Rejected answer 🙅: generated by GPT-2 (openai-community/gpt2) w/ high temperature
    - Chosen answer 🙆: ensembled from textual label

prompt string · *lengths* 1.4k 4.12k	rejected string · *lengths* 31 73	chosen string · *classes* 6 values
You are a medical doctor specialized in diabetes diagnosis. From the provided document, assert if the…	proximal phalanx of the proximal phalan	{"diabetes": "YES"}
You are a medical doctor specialized in diabetes diagnosis. From the provided document, assert if the…	enocarcinoma with high serum hepatitis.	{"diabetes": "YES"}
You are a medical doctor specialized in cancer diagnosis. From the provided document, assert if the…	ography revealed multiple myeloma and right ophthalmic	{"cancer": "YES"}
You are a medical doctor specialized in cancer diagnosis. From the provided document, assert if the…	tomography. The patient had a history of multiple lung	{"cancer": "NO"}

  - Resulting reward model: hanyinwang/layer-project-reward-model
- Update policy model with PPO
  - Training data: samples without binary label data from provided dataset (n = 3,900, a random sample of 200 was used)
  - Resulting model: hanyinwang/layer-project-diagnostic-mistral

RLHF Fine-tuning

- **Fine-tuned Mistral**
  - Prompt ➡️
  - “Diagnostic Mistral”
- **Evaluation**
  - Same eval data & scope as GPT-4
    - 0 if not parseable
  - Not as good as GPT-4
    - Hundred-time size difference
  - Improvement upon fine-tuning 🌟
    - Cancer – 12 misclassified cases
      - Only 1 false positive
      - 3 of the false negatives agreed with GPT-4 output (3146, 2117, and 2840)
    - Diabetes – 2 misclassified cases 👌
      - Both false positives

<u>Mistral prompt template:</u>
*<s>[INST] You are a medical doctor specialized in <condition> diagnosis. From the provided document, assert if the patient **historically and currently** has <condition>. For each condition, only pick from "YES", "NO", or "MAYBE". And you must follow the format without anything further. The results have to be directly parseable with python json.loads().*
Sample output: {"<condition>": "MAYBE"}
Never output anything beyond the format. [/INST]
Provided document: <note>

Table 5. Diagnostic-mistral-7B performances before and after fine-tuning

cancer		accuracy	precision	recall
before **RLHF fine-tune**	**0**	0.60	0.60	1.00
	1		0.00	0.00
after **RLHF fine-tune**	**0**	0.76	0.72	0.97
	1		0.90	0.45

diabetes		accuracy	precision	recall
before **RLHF fine-tune**	**0**	0.90	0.90	1.00
	1		0.00	0.00
after **RLHF fine-tune**	**0**	0.96	1.00	0.96
	1		0.71	1.00

RLHF Fine-tuning – Error analysis 🧐

Table 6B. Error analysis of cancer/diabetes classification using diagnostic-mistral-7B.

ID	diagnostic-mistral response	related snippet from note	analysis	conclusion
2776	\n\n{"cancer": "YES"}</s> False positive	... and slightly **elevated CA-125** and CA-19.9. CT scan...	This patient has a large ovarian mass, for which histology showed a benign mucinous cystadenoma. What might be misleading is the elevated CA-125, which could also be associated with ovarian cancer.	diagnostic-mistral is **wrong**
213	\n\n{"diabetes": "YES"} False positive	...pituitary hormone insufficiency of all anterior axes and **diabetes insipidus** was diagnosed...	"diabetes" was mentioned, but "diabetes insipidus" diagnosed instead of "diabetes mellitus".	diagnostic-mistral is **wrong**
2097	\n\n{"diabetes": "YES"} False positive	Discharge Diagnosis: 1. Congenital perineal groove 2. Neonate of **diabetic mother**	The patient is a neonate with a diabetic mother.	diagnostic-mistral is **wrong**

RLHF Fine-tuning – Error analysis 🧐

Table 6B. Error analysis of cancer/diabetes classification using diagnostic-mistral-7B.

ID	diagnostic-mistral response	related snippet from note	analysis	conclusion
2275	\n\n{"cancer": "MAYBE"} False negative	<none>	The original case presents a rare case of TB infection after liver transplantation. In the original case "His clinical history was remarkable for hepatitis B (HBV) and Genotype 3 hepatitis C (HCV) co-infection, which led to OLT due to hepatocellular carcinoma (HCC),"	diagnostic-mistral is **wrong**, **but corresponding content was not presented in the note**
1523	\n\n{"cancer": ["YES", " False negative	--	the answer was right but messy format	unparseable output
1221	At present, there is no evidence of disease.\n False negative	"... Patient 2, a 32-year-old male with a history of CDH1 mutation and **HDGC**, ..." (hereditary diffuse gastric cancer)	overlook on medical history or insufficient knowledge on abbreviations. Model response resembles original text "At the time of this report, patient 2 is well with no evidence of disease."	diagnostic-mistral is **wrong**
2260	\n\n{"cancer_history": "YES False negative	--	the answer was right but messy format	unparseable output
2762	\n\n{"cancer": "YES", " False negative	--	the answer was right but messy format	unparseable output
1047	\n\n{"prostate_cancer": "Y False negative	--	the answer was right but messy format	unparseable output
1809	\n\n{"cancer": "YES", " False negative	--	the answer was right but messy format	unparseable output
2644	\n\n{"patient_A": {"history": [" False negative	--	we are not sure if the answer is correct. However, this is a multi-patient cases, where two cases were presented. But both cases have cancer.	unparseable output

RLHF Fine-tune -- Summary

- **Performance improvement after fine-tune (even if only 1 epoch on the part of the data)** 🌟
    - Not as good as GPT-4, but there's a hundred-time size difference
    - GPT-4 is not HIPAA complaint, but diagnostic-mistrial is locally deployable
    - Promising if trained further and on more comprehensive data
- **Proposed improvements**
    - Prompt engineering: "diabetes mellitus" v.s. "diabetes", abbreviations
    - SFT in the first step
        - Performance: diabetes > cancer – diabetes is associated a relatively fixed vocabulary, whereas cancer vocab is much bigger w/ various abbr. associated
        - Context is essential
    - Additional data for reward modeling, PPO training, and evaluation
        - Answer pairs: over-simplified – hinders the power of reward model
        - Unlabeled data: subset (200/3,900) used – model under-trained
        - Evaluate on additional conditions -- transferability
        - Evaluate on external datasets – generalizability
    - Enforce output format, e.g., LMFE
    - Solutions outside LLM
        - Classic end-to-end supervised learning
        - Active learning if annotation expertise available
        - Pseudo-labeling / co-training for partially labeled data

Resources

- **GitHub repo:** https://github.com/HanyinWang/layer-project-IMO
- **Reward model:** hanyinwang/layer-project-reward-model
- **Diagnostic-mistral:** hanyinwang/layer-project-diagnostic-mistral
- **Answer pairs for reward model training:** hanyinwang/layer-project-reward-training