

# **Cancer and diabetes patient phenotyping**

-- A large language model (LLM) solution

Hanyin Wang

May 15, 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 6. 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 Reinforcement 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](#)

## 3. Retrieval Augmented Generation (RAG)

# Zero-shot GPT4




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

Table 7. Answer-label correspondence.

answer	label
YES	1
NO	0
MAYBE	

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

## **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 6. 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)

Table 8. 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 9. 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 <b>chronic lymphocytic leukemia</b> (in remission). ...	The case was not focused on cancer, but the patient did have a medical history of leukemia.	GPT is <b>wrong</b>
3146	{"cancer": "MAYBE"}	...The mass was diagnosed as <b>LCH</b> . ...	In the original case report, the following points were made: "Lobular capillary hemangioma is a rare, rapidly growing, <b>benign tumor</b> ", meaning that LCH is not considered as cancer. However, this information was not included in the note. Furthermore, according to <a href="#">NIH</a> , "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 <b>right</b>
2117	{"cancer": "NO"}	... Tumor histology was reminiscent of <b>desmoid fibromatosis</b> and consistent with <b>desmoplastic fibroma</b> . ...	Both are <b>benign</b> 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 <b>right</b>
2840	{"cancer": "NO"}	... "The primary tumor was diagnosed as a <b>glomus tumor</b> based on these findings. " ...	The glomus tumors reported have been mostly benign neoplasms and <b>very rarely malignant</b> . 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 <b>basis for the lesion being benign</b> . ..."	The conclusion of GPT is <b>correct</b> , 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

- 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	rejected string · lengths	chosen string · classes
		
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](#)

Table 6. 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)

# 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

**Mistral prompt template:**

*<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 10. Diagnostic-mistral-7B performances before and after fine-tuning*

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

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



# RLHF Fine-tuning – Error analysis 🤔

Table 11B. Error analysis of cancer/diabetes classification using diagnostic-mistral-7B.

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

# RLHF Fine-tuning – Error analysis 🤔

Table 11B. 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 <u>hepatocellular carcinoma (HCC)</u> ,"	diagnostic-mistral is <b>wrong</b> , but <b>corresponding content was not presented in the note</b>
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 <b>HDGC</b> , ..." (hereditary diffuse gastric cancer)	overlook on medical <u>history</u> or insufficient knowledge on <u>abbreviations</u> . Model response <u>resembles original text</u> "At the time of this report, patient 2 is well with no evidence of disease."	diagnostic-mistral is <b>wrong</b>
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 compliant, but diagnostic-mistral 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

# Improvement -- RAG

- **Further performance improvement for cancer** (5 false negatives)
  - 3146, 2840 align with GPT-4 (possible wrong label)
  - 2275 (possible wrong label), 2644 (multi-case) analyzed earlier
  - New case: 710
- **Even performance for diabetes** (1 FN & 1 FP)
  - 2097 (diabetes insipidus)
  - New case: 2762

cancer		accuracy	precision	recall
before RLHF	0	0.60	0.60	1.00
	1		0.00	0.00
after RLHF	0	0.76	0.72	0.97
	1		0.90	0.45
after RLHF & RAG	0	0.90	0.86	1.00
	1		1.00	0.75

diabetes		accuracy	precision	recall
before RLHF	0	0.90	0.90	1.00
	1		0.00	0.00
after RLHF	0	0.96	1.00	0.96
	1		0.71	1.00
after RLHF & RAG	0	0.96	0.98	0.98
	1		0.80	0.80

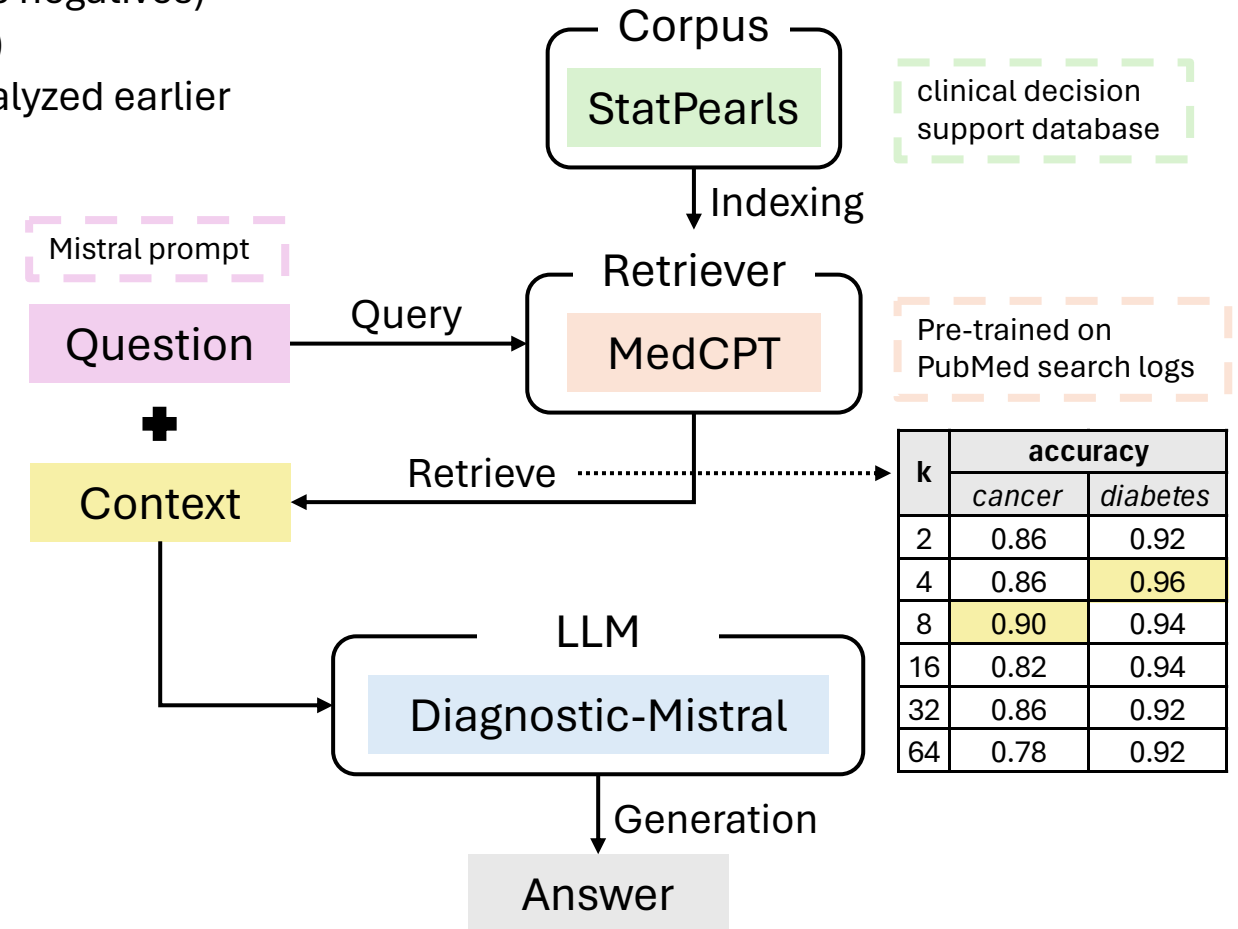


Figure 11. RAG pipeline

# RLHF Fine-tuning + RAG – Error analysis 🤔

Table 12. Error analysis of cancer/diabetes classification using diagnostic-mistral-7B + RAG.

ID	diagnostic-mistral + RAG response	related snippet from note	analysis	conclusion
710	<code>{"cancer": "NO"}</code> False negative	The patient, a 54-year-old man diagnosed with <b>pulmonary large-cell neuroendocrine carcinoma</b> , .... responded well to treatment with metilprednisolone and piperaciline-tazobactam for <b>pneumonitis</b> .	A pulmonary cancer patient with pneumonitis. Retrieved content titles: 1. 'Cryptogenic Organizing Pneumonia -- Treatment / Management' 2. 'Nonspecific Interstitial Pneumonia -- Treatment / Management' 3. 'Farmer's Lung -- Treatment / Management' 4. 'Pneumonia Pathology -- Treatment / Management'	diagnostic-mistral (RAG) was <b>wrong</b> . Retrival content focused only on pneomonia and ignored cancer.
2762	<code>{"diabetes": "NO"}</code> False negative	The patient had a known history of <b>diabetes mellitus</b> , hypertension, and coronary artery disease.	This patient was diagnosed with <b>thyroid carcinoma</b> . Retrieved content titles: 1. 'Thyroidectomy -- Indications -- Initial Evaluation' 2. 'Iodine-131 Uptake Study -- Potential Diagnosis' 3. 'Substernal Thyroidectomy -- Indications' 4. 'Follicular Adenoma -- Treatment / Management -- Medical Management'	diagnostic-mistral (RAG) was <b>wrong</b> . Retrival content diluted useful inforamtion in the medical history section.

# RLHF Fine-tuning + RAG – Improvement

Table 13. Strength analysis of cancer/diabetes classification using diagnostic-mistral-7B + RAG.

ID	diagnostic-mistral + RAG response	related snippet from note	analysis	conclusion
2776	{"cancer": "NO"} True negative	Histology of the cyst showed a <b>benign mucinous cystadenoma</b> .	Large ovarian mass, benign 'Retrieved content titles: 1. 'Tubo-Ovarian Abscess -- Differential Diagnosis' 2. 'Acute Abdomen -- Consultations' 3. 'Postoperative Fever -- History and Physical -- Subacute Fever' 4. 'Pelvic Abscess -- Introduction'	RAG helped focus on major diagnosis.
213	{"diabetes": "NO"} True negative	..., where pituitary hormone insufficiency of all anterior axes and <b>diabetes insipidus</b> was diagnosed, and treatment with various hormones was implemented. Final histological diagnosis revealed <b>autoimmune hypophysitis</b> .	"autoimmune hypophysitis" aka "lymphocytic hypophysitis" Retrieved content titles: 'Lymphocytic Hypophysitis -- Treatment / Management'  RAG helped focus on major diagnosis, we still need to emphasize on `historical` and `current`	RAG improved the performance by dragging more attention to the current diagnosis.

# RLHF Fine-tuning + RAG – additional analysis on RAG

- **Analysis on medical term abbreviations**
  - HDGC-related concept appears in `Gastric cancer -- Etiology` -- was not retrieved

*Table 14. Additional analysis on RAG of cancer/diabetes classification using diagnostic-mistral-7B + RAG.*

ID	diagnostic-mistral response	related snippet from note	analysis	conclusion
1221 no RAG	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 <b>HDGC</b> , ..." (hereditary diffuse gastric cancer)	overlook on medical <u>history</u> or insufficient knowledge on <u>abbreviations</u> . Model response <u>resembles original text</u> "At the time of this report, patient 2 is well with no evidence of disease."	diagnostic-mistral is <b>wrong</b>
1221 RAG	{"cancer": "YES"}		Retrieved content titles: 1. 'Acute Pancreatitis -- Consultations' 2. 'Pancreaticoduodenectomy -- Complications' 3. 'Abdominal Angina -- Pearls and Other Issues' 4. 'Acute Pancreatitis -- Etiology'	diagnostic-mistral + RAG is <b>correct</b> but was not able to retrieve anything directly related to HDGC

# RLHF Fine-tune + RAG -- Summary

- **Performance improvement with RAG 🏆**
  - RAG increases attention on major diagnosis
  - Ignorance on medical history is exaggerated (the thyroid cancer example)
  - Promising if the model undergone SFT and if RLHF trained further
- **Proposed improvements**
  - Find proper `k` on a validation set
  - Indexing and vector database
    - Current implementation: Flat IP with FAISS, ~3 it/s
    - Searching speed and memory issue
    - Structured / hierarchical index (e.g. IVF, HNSW)
  - Corpus content
    - Still not comprehensive enough, especially in abbreviations (the HDGC example)
  - Retriever – MedCPT
    - Trained on PubMed, not completely clinical decision support
    - Better encoder on more relevant pre-pretraining corpora



# Resources

- GitHub repo: <https://github.com/HanyinWang/layer-project-IMO>
- Reward model: [hanyinwang/layer-project-reward-model](https://github.com/HanyinWang/layer-project-reward-model)
- Diagnostic-mistral: [hanyinwang/layer-project-diagnostic-mistral](https://github.com/HanyinWang/layer-project-diagnostic-mistral)
- Answer pairs for reward model training: [hanyinwang/layer-project-reward-training](https://github.com/HanyinWang/layer-project-reward-training)