Counting Clues:
A Lightweight Probabilistic Baseline Can Match an LLM

Anonymous ACL submission

Abstract

Large language models (LLMs) excel on multiple-choice clinical diagnosis benchmarks, yet it is unclear how much of this performance reflects underlying probabilistic reasoning. We study this through questions from MedQA, where the task is to select the most likely diagnosis. We introduce the Frequency-Based Probabilistic Ranker (FBPR), a lightweight method that scores options with a smoothed Naive Bayes over concept-diagnosis co-occurrence statistics from a large corpus. When co-occurrence statistics were sourced from the pretraining corpora for OLMo and Llama, FBPR achieves comparable performance to the corresponding LLMs pretrained on that same corpus. Direct LLM inference and FBPR largely get different questions correct, with an overlap only slightly above random chance, indicating complementary strengths of each method. These findings highlight the continued value of explicit probabilistic baselines: they provide a meaningful performance reference point and a complementary signal for potential hybridization. While the performance of LLMs seems to be driven by a mechanism other than simple frequency aggregation, we show that an approach similar to the historically grounded, low-complexity expert systems still accounts for a substantial portion of benchmark performance.

1 Introduction

Large Language models (LLMs) have been used for tasks requiring probabilistic reasoning, both direct inference and estimation of priors (Paruchuri et al., 2024; Feng et al., 2024; Nafar et al., 2025). However, our understanding of probabilistic reasoning in LLMs is still nascent. In this work, we study the probabilistic reasoning required to choose the most likely medical diagnosis given a clinical scenario. LLMs have achieved high performance across medical benchmarks, leading to speculation that they possess advanced clinical reasoning capabilities (Kung et al., 2022; Singhal et al., 2022, 2025; Nori et al., 2023). This narrative is reinforced by claims of LLMs achieving diagnostic performance on complex cases that surpass human experts (McDuff et al., 2025; Nori et al., 2025). However, a growing number of works also reveal potential deficiencies in the clinical reasoning and decision-making of current LLMs(Chen et al., 2024). For example, expert systems such as DxPlain still surpass LLMs on unpublished clinical cases for diagnoses (Barnett et al., 1987; Berner et al., 1994; Feldman et al., 2025). This suggests there may be a disconnect between performance on medical licensing exams and true probabilistic reasoning, with research highlighting LLM failures in meta-cognition (Griot et al., 2025a), processing new information under uncertainty (McCoy et al., 2025), and aligning with the reasoning process of physicians on medical question-answering (Hao et al., 2025).

Building on these observations, a key evaluation gap is clarifying how much reported accuracy on medical QA benchmarks reflects exploitation of corpus-level co-occurrence statistics versus more structured clinical probabilistic reasoning. Widely used biomedical QA benchmarks are predominantly knowledge-heavy (67.2% of questions across 11 datasets), with models performing worse on the reasoning-heavy subsets (Thapa et al., 2025). Complementary synthetic evidence shows that LLMs can achieve 64% on a fictional benchmark about an invented organ (Glianorex), while physicians reach only 27% (Griot et al., 2025b), indicating that strong scores can arise from linguistic cues and test-taking heuristics over domain knowledge. These patterns suggest revisiting earlier paradigms that leveraged statistical association, such as earlier clinical decision-support systems (CDSS) like DxPlain (Barnett et al., 1987; Berner et al., 1994). DxPlain combines computer-interpretable clinical guidelines with an expert-

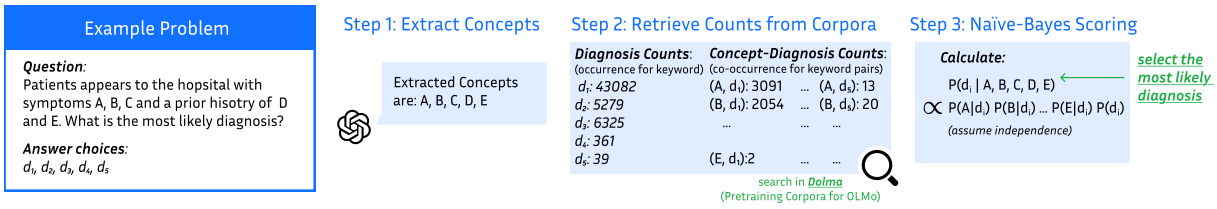

Figure 1: Frequency-Based Probabilistic Ranker: (1) concept extraction from the question, (2) corpus frequency retrieval, and (3) calculate a Naive-Bayes scoring to select the most likely diagnosis.

designed, data-refinable Bayesian network to provide probabilistic, patient-specific decision support. As previously mentioned, it could even surpass LLMs on unpublished clinical cases (Nee and Hein, 2010; Elkin et al., 2010; Bauer et al., 2002; Feldman et al., 2025).

Beyond medicine, recent work leverages LLM-derived probabilities for Bayesian reasoning, either aligning Bayesian networks with LLM-abduced factors (BIRD) (Feng et al., 2024) or eliciting conditional probabilities to parameterize Bayesian Networks (Nafar et al., 2025).Therefore, here we would like to investigate how much a coarser, low-complexity, and transparent, frequency-based method applied directly to an LLM's pretraining corpus can achieve on diagnosis questions. Therefore, we introduce a simple **Frequency-Based Probabilistic Ranker (FBPR)** that eschews complex reasoning (Figure 1). Our method models diagnosis using a Naive Bayes-like probability derived from the co-occurrence frequency of diagnosis and clinical concepts from large-scale pretraining corpora, Dolma (Soldaini et al., 2024) and RedPajama (Weber et al., 2024). We evaluate FBPR on a diagnostic subset of MEDQA-USMLE, a dataset that consists of United States Medical Licensing Examination-style multiple-choice questions (Jin et al., 2021). We also evaluate OLMo Instruct 7B(Groeneveld et al., 2024), an open-weight instruction-finetuned LLM pretrained on the Dolma corpus, and LLaMA 65B(Touvron et al., 2023a), which is pretrained on the RedPajama corpus, for performance comparison with FBPR applied on their respective pretraining corpora. We found that Dolma-based FBPR achieved an accuracy (46.7%) substantially above random chance (20%) and closely matched OLMo Instruct (44.1%). Similar results were obtained from FBPR based on RedPajama (44.5%) and LLaMA 65B (47.0%). Predictions from FBPR and LLMs are found to be complementary, highlighting the potential benefits of combining simple statistical and LLM-based approaches.

2 Methodology

2.1 Data and Task Formulation

We evaluate on two transparent corpus-model pairs: (1) Dolma and OLMo Instruct 7B, and (2) RedPajama and LLaMA 65B. OLMo Instruct 7B is pretrained on the Dolma corpus and further instruction-tuned for instruction following abilities. LLaMA-65B is a substantially larger LLM pretrained on RedPajama and shows usable instruction-following on MedQA. This transparency enables a fair, corpus-aligned comparison between our statistical method and the LLM baselines.

We focus on a diagnosis subset of the MEDQA benchmark ($n = 719$; details in Appendix A), which was not included in Dolma, instruction-tuning dataset for OLMo Instruct, or RedPajama. Each question consists of a patient scenario followed by a prompt for the most likely diagnosis and five candidate diagnoses $D = \{d_1, d_2, d_3, d_4, d_5\}$. We use the five-option version to match the setting of the actual US medical licensing exam, and the goal is to predict the most likely diagnosis $d^\star \in D$ from the scenario description.

Out of these 719 questions, 256 are from the Step 1 domain of the US medical licensing exam, which focuses on basic science knowledge, and 463 are from the Step 2&3, which emphasize clinical reasoning and patient management. It is widely recognized that, while Step 1 emphasizes factual recall, Steps 2&3 require an application of knowledge through clinical reasoning and therefore demand stronger reasoning ability (Jin et al., 2021).

2.2 Frequency-Based Probabilistic Ranker

Our pipeline proceeds through the three stages depicted in Fig. 1 and detailed below:

Stage 1: concept extraction For each question, we prompt the GPT-4o to extract k concise (each ≤ 4 words) clinical concepts $\mathbf{x} = \{x_1, \ldots, x_k\}$ that were emphasized in the presented clinical scenario. The prompt omits any mention of diagnosis to avoid bias toward selectively "diagnostic" fea-

tures and instead aims to obtain an objective set of symptoms mentioned. We further ask GPT-4o to label each extracted concept with whether it is mentioned positively or negated. The prompt can be found in Appendix B, with examples of extracted concepts in Appendix F.3.

Stage 2: corpus frequency retrieval We leverage *Infini-gram* (Liu et al., 2024), an efficient tool for querying token and phrase frequencies in massive corpora, to obtain from Dolma or RedPajama

- the number of occurrences of candidate diagnosis d as $C(d)$, and
- the number of co-occurrences of candidate diagnosis d along with extracted concept x_i as $C(d, x_i)$.

We also expand each candidate diagnosis and extracted concept by their variations in spacing and capitalized version to account for the fact that *Infini-gram* performs tokenization before searching. More details are in Appendix C.

Stage 3: Naive-Bayes scoring Assuming concept independence and their conditional independence given each candidate diagnosis, we estimate the posterior $P(d \mid x_1, \ldots, x_k)$ for each candidate diagnosis and choose the one with the highest probability given the extracted concepts.

Let $d \in D$ and abbreviate $\mathbf{x} = \{x_1, \ldots, x_k\}$.

$$P(d \mid \mathbf{x}) = \frac{P(\mathbf{x} \mid d)\, P(d)}{P(\mathbf{x})} \tag{1}$$

$$= \frac{\left[\prod_{i=1}^{k'} P(x_i \mid d)\right] P(d)}{P(\mathbf{x})} \tag{2}$$

$$\approx \frac{\left[\prod_{i=1}^{k'} \frac{C'(d,x_i)}{C(d)}\right] \frac{C(d)}{N}}{P(\mathbf{x})} \tag{3}$$

$$= \frac{C(d)^{-(k'-1)}}{N \cdot P(\mathbf{x})} \left[\prod_{i=1}^{k'} C'(d, x_i)\right]. \tag{4}$$

Step (2) is based on the naive conditional-independence assumption. Step (3) substitutes empirical estimates $P(x_i \mid d) \approx C'(d, x_i)/C(d)$ and $P(d) \approx C(d)/N$, where N is the total token count of the pre-training corpus.

Given the affirmation/negation label, we took $C'(d, x_i) = C(d, x_i)$ for every (candidate diagnosis, affirmed concept) pair and excluded the negated concepts from the score calculation. k' thus equaled the number of affirmed concepts in each question.

The denominator $N \cdot P(\mathbf{x})$ does not depend on d and is dropped in the calculation of the score for comparing candidate diagnoses. With Laplace smoothing ($\delta > 0$) and a logarithm, we have the score for each candidate diagnosis as

$$S_d = \sum_{i=1}^{k'} \log\left(C'(d, x_i) + \delta\right) - (k' - 1) \log\left(C(d) + \delta\right)$$

and the answer by FBPR is the candidate diagnosis with the highest score within each question:

$$\hat{d} = \text{argmax}_{d \in D} S_d.$$

3 Results

3.1 FBPR Performance

Method	Accuracy
Random Chance	20%
FBPR on Dolma	46.7%
OLMo Instruct 7B	44.1%
FBPR on RedPajama	44.5%
LLaMA 65B	47.0%

Table 1: Prediction accuracy on five-option MEDQA diagnosis subset (n = 719).

***k*=5** We prompted GPT-4o to identify five concepts from each MEDQA question and assign each concept an affirmation or negation label. 16 of the 719 questions contained negated concepts (details see Appendix B).

With k=5, Dolma-based FBPR ranked the correct diagnosis highest ($\hat{d} = d^\star$, i.e., made correct predictions) in 336 questions, corresponding to an accuracy of 46.7% while RedPajama-based FBPR achieved 44.5% accuracy (Table 1). The correct diagnosis is most frequently assigned the highest rank by FBPR, with progressively fewer instances receiving ranks 2 through 5 (Figure 2). The observed monotonic pattern indicates that this simple frequency-based method appears to capture clinically meaningful links between presenting concepts and diagnoses. The same pattern holds for RedPajama-based FBPR (Figure 5.

Dolma-based FBPR correctly identified the most likely diagnosis for 51% of Step 1 questions but only 44.3% of Step 2&3 questions (Table 2 in Appendix). This difference likely reflects the heavier multi-step reasoning demands of Step 2&3, where FBPR's occurrence frequency-based reasoning is less effective. However, the gap is modest, suggesting that even simple frequency-based Bayesian

strategies can account for a notable portion of performance often attributed to complex reasoning.

We further explored the relationship between FBPR's internal certainty and its actual correctness by using the top-1 scores softmaxed within each question as a proxy for certainty (Figure 6 in Appendix). For Dolma-based FBPR, correctly predicted questions have higher softmaxed top scores compared to incorrectly predicted ones (median 0.94 vs. 0.86). 27.4% of questions are with a softmaxed highest score exceeding 0.99; the accuracy of FBPR on these questions is 59.4%, higher than the general 46.7% achieved on the full dataset.

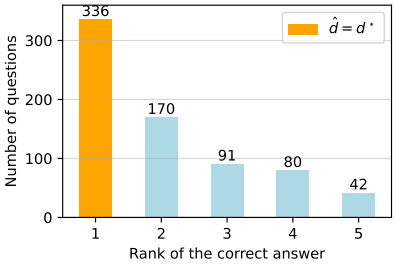

Figure 2: Distribution of the rank of the correct diagnosis of each question by Dolma-based FBPR with $k = 5$.

Unrestricted *k* We also experimented prompting GPT-4o without a restriction on k (details see Appendix B). Under this setting, FBPR on RedPajama achieved an accuracy of 40.9%. Its rank distribution follows the same decreasing pattern as under $k = 5$ (Figure 5 in Appendix) and has a slightly higher performance on Step 2&3 than Step 1 (Table 2 in Appendix).

3.2 Alignment with LLM Performance

To assess alignment with a model trained on the same underlying distribution, we compare the Dolma-based FBPR to OLMo Instruct 7B (pretrained on the Dolma corpora) and the RedPajam-based FBPR to LLaMA 65B, both under a zero-shot setting (details in Appendix D).

Dolma-based FBPR and OLMo Instruct are both correct on 24.8% of questions. This observed joint accuracy is slightly higher than the expected joint success of around $44.1\% \times 46.7\% \approx 20.6\%$ under an independence assumption. This indicates that the methods are only mildly correlated. FBPR alone is correct on 21.9% of questions, and OLMo alone on 19.3%.

Across the diagnostic subset, the two methods give exactly the same answer to 38.3% of the questions (shown as the diagonal in Figure 3). 64.7% of these agreed answers are correct. This indicates that inter-method agreement is a helpful reliability cue. The same pattern appeared in the comparison of RedPajama-based FBPR and LLaMA 65B (Appendix F.1).

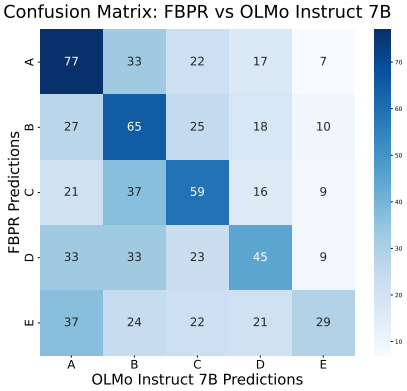

Figure 3: Confusion matrix between the Dolma-based FBPR and OLMo Instruct 7B predictions on MEDQA diagnosis subset (n=719).

While the performance of LLMs appears to be driven by mechanisms beyond simple frequency aggregation, we show that a historically grounded, low-complexity expert-system–style method based on their pretraining corpora statistics accounts for a substantial portion of benchmark performance. As the two methods' signals only partially overlap, FBPR has the potential to offer complementary diagnostic information to LLMs, motivating hybrid approaches to provide better diagnosis, robustness, and more evidence-supported decision-making.

4 Dicussion and Future Work

Our simple frequency-based concept–diagnosis co-occurrence method attains accuracy on MEDQA diagnostic questions comparable to LLMs pretrained on the same corpora, indicating that surface distributional priors can account for a substantial portion of benchmark performance. Despite similar accuracy, the LLM appears to exploit different mechanisms over the same corpus and thus offers complementary signals.

Our baseline in this work is quite naive: it performs no synonym or semantic normalization (e.g., "high blood pressure," "elevated blood pressure," and "hypertension" remain separate), which could further improve its scores. Next steps also include developing probes that can better distinguish memorized pattern retrieval from genuine reasoning.

Limitations

Our analysis targets a diagnosis-focused subset of MEDQA, and the concept set is extracted by an external LLM, making results sensitive to prompting and extraction choices. Frequency estimates are obtained via *Infini-gram* (Liu et al., 2024), which returns occurrence-level (rather than document-level) counts and approximates very frequent strings; precise document-level co-occurrence is deferred due to resource constraints. We further limit evaluation to two transparent corpus–model pairs: (1) Dolma and OLMo Instruct 7B, (2) RedPajama and LLaMA 65B. Neither is state-of-the-art on this benchmark and thus may not reflect the frontier of LLM performance. Nonetheless, our aim is not to introduce a high-performing system but to isolate how far an explicit, low-complexity probabilistic baseline can go using pretraining corpus statistics.

Ethics Statement

In writing this paper, we used an AI assistant to correct grammatical errors. During the coding process, we utilized AI tools for code completion.

References

G Octo Barnett, James J Cimino, Jon A Hupp, and Edward P Hoffer. 1987. Dxplain: an evolving diagnostic decision-support system. *Jama*, 258(1):67–74.

Brent A Bauer, Mark Lee, Larry Bergstrom, Dietlind L Wahner-Roedler, John Bundrick, Scott Litin, Edward Hoffer, Richard J Kim, Kathleen Famiglietti, G Octo Barnett, and 1 others. 2002. Internal medicine resident satisfaction with a diagnostic decision support system (dxplain) introduced on a teaching hospital service. In *Proceedings of the AMIA Symposium*, page 31.

Eta S Berner, George D Webster, Alwyn A Shugerman, James R Jackson, James Algina, Alfred L Baker, Eugene V Ball, C Glenn Cobbs, Vincent W Dennis, Eugene P Frenkel, and 1 others. 1994. Performance of four computer-based diagnostic systems. *New England Journal of Medicine*, 330(25):1792–1796.

Canyu Chen, Jian Yu, Shan Chen, Che Liu, Zhongwei Wan, Danielle Bitterman, Fei Wang, and Kai Shu. 2024. Clinicalbench: Can llms beat traditional ml models in clinical prediction? *arXiv preprint arXiv:2411.06469*.

Peter L Elkin, Mark Liebow, Brent A Bauer, Swarna Chaliki, Dietlind Wahner-Roedler, John Bundrick, Mark Lee, Steven H Brown, David Froehling, Kent Bailey, and 1 others. 2010. The introduction of a diagnostic decision support system (dxplain™) into the workflow of a teaching hospital service can decrease the cost of service for diagnostically challenging diagnostic related groups (drgs). *International journal of medical informatics*, 79(11):772–777.

Mitchell J Feldman, Edward P Hoffer, Jared J Conley, Jaime Chang, Jeanhee A Chung, Michael C Jernigan, William T Lester, Zachary H Strasser, and Henry C Chueh. 2025. Dedicated ai expert system vs generative ai with large language model for clinical diagnoses. *JAMA Network Open*, 8(5):e2512994–e2512994.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. 2024. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*.

Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025a. Large language models lack essential metacognition for reliable medical reasoning. *Nature communications*, 16(1):642.

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025b. Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5321–5341.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya

Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Yuexing Hao, Kumail Alhamoud, Hyewon Jeong, Haoran Zhang, Isha Puri, Philip Torr, Mike Schaekermann, Ariel D Stern, and Marzyeh Ghassemi. 2025. Medpair: Measuring physicians and ai relevance alignment in medical question answering. *arXiv preprint arXiv:2505.24040*.

Daniel P Jeong, Saurabh Garg, Zachary C Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? *arXiv preprint arXiv:2411.04118*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. 2022. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Liam G McCoy, Rajiv Swamy, Nidhish Sagar, Minjia Wang, James Cao, Stephen Bacchi, Nigel Fong, Nigel CK Tan, Kevin Tan, Thomas A Buckley, and 1 others. 2025. Do language models think like doctors? *medRxiv*, pages 2025–02.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7.

Aliakbar Nafar, Kristen Brent Venable, Zijun Cui, and Parisa Kordjamshidi. 2025. Extracting probabilistic knowledge from large language models for bayesian network parameterization. *arXiv preprint arXiv:2505.15918*.

Oliver Nee and Andreas Hein. 2010. Clinical decision support with guidelines and bayesian networks. *Decision Support Systems, Advances in, Book, editor Ger Devlin*, pages 117–137.

Harsha Nori, Mayank Daswani, Christopher Kelly, Scott Lundberg, Marco Tulio Ribeiro, Marc Wilson, Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren, and 1 others. 2025. Sequential diagnosis with language models. *arXiv preprint arXiv:2506.22405*.

Harsha Nori, Nicholas King, S. McKinney, Dean Carignan, and E. Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *ArXiv*, abs/2303.13375.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*.

K. Singhal, Shekoofeh Azizi, T. Tu, S. Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, Martin G. Seneviratne, P. Gamble, C. Kelly, Nathaneal Scharli, A. Chowdhery, P. A. Mansfield, B. A. Y. Arcas, D. Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Nature*, 620:172 – 180.

Karan Singhal, Tao Tu, Juraj Gottweis, R. Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Q. Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, P. Mansfield, and 16 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31:943 – 950.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, and 17 others. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, and 1 others. 2025. Disentangling reasoning and knowledge in medical large language models. *arXiv preprint arXiv:2505.11462*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, and 1 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maurice Weber, Daniel Y. Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, Ben Athiwaratkun, Rahul Chalamala, Kezhen Chen, Max Ryabinin, Tri Dao, Percy Liang, Christopher Ré, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. *NeurIPS Datasets and Benchmarks Track*.

A Diagnosis subset formation

The MEDQA benchmark is a collection of United States Medical Licensing Examination (USMLE)-style questions. Each question stem consists of a clinical vignette describing a patient scenario, followed by a query that typically requires identifying the most likely diagnosis, the appropriate next step in management, or a related clinical decision. We used the version where each question provides five candidate answer options, exactly one of which is correct.

We focus on diagnosis-related questions in the MEDQA training set. We first identified questions in which the query (the last sentence of the question stem) explicitly contains the word "diagnosis". After converting all query text to lowercase, we found and kept questions with the top-two most common phrasings: "which of the following is the most likely diagnosis?" (588 questions) and "what is the most likely diagnosis?" (131 questions). This query sentence in each question was stripped before feeding the clinical scenario into GPT-4o for concept extraction and labeling.

B Concept extraction and labeling

For both concept extraction and affirmation/negation labeling, we queried the GPT-4o API with `temperature` set to 0.0 and `top_logprobs` set to 0, ensuring mostly deterministic outputs. Example extracted concepts and labels can be found in Appendix F.3.

B.1 *k*=5

We prompted GPT-4o for concept extraction with a system prompt

> "You are a medical expert that extracts the five most informative clinical concepts from each passage."

and a query prompt encodes both the number of concepts to extract (`n`=5) and the question stem (`cleaned_question`), excluding the trailing query such as "What is the most likely diagnosis?".

> Given the passage below, list exactly {`n`} canonical **clinical** concepts.
>
> Rules for each concept:
> • Must be clinically relevant (symptom, physical sign, diagnosis, or treatment).
> • Prefer the passage's wording **if** it is already $\leq$ 4 words; otherwise shorten or normalize to a standard term $\leq$ 4 words (no numbers or strength grades).
> • Skip generic exam words and any term containing the stems: exam, test, lab, imaging, manage, work-up.
> • The term itself must not contain a comma.
>
> Output formatting: lower-case, comma-separated list of exactly {`n`} concepts, with no other text.
>
> Passage:
> {`cleaned_question`}

Then the extracted concepts (`keywords`) along with the cleaned question stem (`question_text`) are sent to GPT-4o for affirmation/negation labeling in a query prompt

> Question:
> `question_text`
>
> Keywords (comma-separated, KEEP ORDER):

```
', '.join(keywords)
```

Return the single required line exactly as specified.

alongside a system prompt

> You are a precise medical NLP assistant tagging keyword polarity for MedQA-style prompts.
>
> Task:
> Given a MedQA question text and an ordered, comma-separated list of keywords, decide for EACH keyword whether it is affirmed/present/tested ("positive") or explicitly denied/absent/ruled-out ("negative") based ONLY on the question text.
>
> Rules:
> - "positive" if the question affirms/treats the keyword as present/relevant/tested or as a correct/true option.
> - "negative" ONLY if the question explicitly negates or rules it out (e.g., "NOT", "EXCEPT", "absent", "no evidence of", "which is NOT").
> - If merely mentioned (no explicit negation), label "positive".
> - Be careful with NOT/EXCEPT questions: items asked for as NOT/EXCEPT are "negative".
> - Judge strictly from the given text; do not add outside knowledge.
>
> Output FORMAT (STRICT):
> Return a SINGLE LINE string containing EXACTLY one item per keyword in the SAME ORDER, separated by comma+space,
> each item in the form "<keyword>: positive" or "<keyword>: negative".
> Do NOT add any extra text, quotes, explanations, or newlines.

16 out of 719 MEDQA diagnosis subset questions contain negated concepts (15 with a single negation and 1 with two).

B.2 No restriction on *k*

We also explored with prompting GPT-4o without restricting k for concept extraction. Specifically, "five" in the concept extraction system prompt and "exactly {n}" in the query prompt were omitted. GPT-4o returned a varying k across questions with a median (IQR) of 9 (7, 11) (Figure 4). 44 questions had negated concepts (30 with a single negated concept, 6 with two, 7 with three, and 1 with five).

C *Infini-gram* usage

Infini-gram provides a large-scale n-gram index that enables fast retrieval of occurrence and co-occurrence statistics from web-scale corpora. It performs search at the token level, converting input words or phrases into tokens before counting their occurrences. Variations such as the presence of a leading space and lowercase versus capitalized forms influence how tokens are generated and therefore affect the returned counts. Our interest, however, lies in the frequency of candidate diagnoses or (candidate diagnosis, concept) pairs in a case-agnostic and position-agnostic manner. To address this, we generated four surface-form variations for each candidate diagnosis or concept (original, leading space, all lowercase,

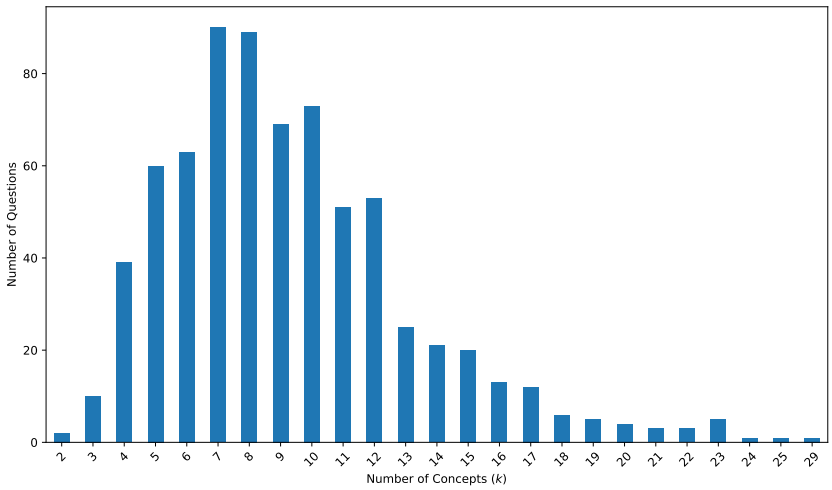

Figure 4: Distribution of the number of concepts per question extracted by GPT-4o when prompt contains no restriction.

and first-letter capitalized) and combined them using an OR clause, i.e.,

`text_original` OR
`text_original_leading_space` OR
`text_all_lowercase` OR
`text_all_lowercase_leading_space` OR
`text_capitalized` OR
`text_capitalized_leading_space`.

OR clauses are supported by *Infini-gram* to obtain aggregated occurrence counts for the OR connected components, and thus help return our desired occurrence frequency of candidate diagnoses.

To obtain co-occurrence of a candidate diagnosis and a concept, we feed *Infini-gram* with a clause connecting their OR clauses with an AND operator, i.e.,

(
`diagnosis_original` OR
`diagnosis_original_leading_space` OR
`diagnosis_all_lowercase` OR
`diagnosis_all_lowercase_leading_space` OR
`diagnosis_capitalized` OR
`diagnosis_capitalized_leading_space`
) AND (
`concept_original` OR
`concept_original_leading_space` OR
`concept_all_lowercase` OR
`concept_all_lowercase_leading_space` OR
`concept_capitalized` OR
`concept_capitalized_leading_space`).

Note that we de-duplicate the text variations before connecting them in OR clause. For most words or phrases, there are four unique variants: all-lowercase with and without a leading space, and initial-capitalized with and without a leading space. Some text, for example the ones with internal capitalization,

have more than four variations created by our algorithm. A related complication is that *Infini-gram* supports OR clauses with at most four components, which requires splitting the query into multiple OR clauses or OR–AND clauses. For the former, a summation of individual OR clause count is sufficient for the overall count; for the latter, we apply inclusion–exclusion to compute the counts.

By *Infini-gram* design, the occurrences of clauses connected by AND can only be examined within a window of 1000 tokens at maximum. When any clause connected by AND has an occurrence frequency larger than 500000, the returned co-occurrence frequency is an estimate.

We used two indices provided by *Infini-gram*:

- `v4_dolma-v1_7_llama`: Dolma-v1.7 corpus tokenized by LLaMA-2 (Touvron et al., 2023b) consists of 3,403,336,408 documents and 2,604,642,372,173 tokens.
- `v4_rpj_llama_s4`: Red Pajama corpus tokenized by LLaMA-2 consists of 931,361,530 documents and 1,385,942,948,192 tokens.

D Large language models evaluation

We performed zero-shot prompting to get LLM prediction of most-likely diagnosis. Below is an example query for LLM:

Question: A 70-year-old man comes to the physician for the evaluation of an 8-week history of blood in his stool. Two months ago, he had an episode of bronchitis and was treated with amoxicillin. Since then, he has noticed blood in his stool and on the toilet paper occasionally. The patient has had intermittent constipation for the past 5 years. Six months ago, he had severe left lower quadrant pain and fever that resolved with antibiotic therapy. He underwent a colonoscopy 3 years ago, which did not show any evidence of malignancy. He takes levothyroxine for hypothyroidism. He had smoked one pack of cigarettes daily for 45 years, but quit smoking 10 years ago. He drinks one glass of red wine every night. He appears pale. He is 180 cm (5 ft 11 in) tall and weighs 98 kg (216 lb); BMI is 32 kg/m2. His temperature is 36°C (96.8°F), pulse is 85/min, and blood pressure is 135/80 mm Hg. Physical examination shows pale conjunctivae. Cardiopulmonary examination shows no abnormalities. The abdomen is soft and nontender with no organomegaly. Digital rectal examination shows no masses. Test of the stool for occult blood is positive. Laboratory studies show:
Hemoglobin 11 g/dL
Mean corpuscular volume 76 μm^3
Red cell distribution width 17% (N = 13–15)
Leukocyte count 5,000/mm^3
Which of the following is the most likely diagnosis?"

Options:
A: Colorectal carcinoma
B: Diverticulosis
C: Ischemic colitis
D: Hemorrhoids
E: Pseudomembranous colitis

Answer:

OLMo_Instruct_0724 generated the following output in response to the query above:

B: Diverticulosis

Explanation:

> The patient's history of intermittent constipation, positive occult blood test, and the absence of any evidence of malignancy on previous colonoscopy make diverticulosis the most likely diagnosis.

An extractor was created to extract the answer from LLM response. In the example above, the extracted answer refers to option choice B. Our extractor operates in three tiers, mirroring the pipeline of Jeong et al. (2024):

1. It first applies a strict regex to find an exact, stand-alone option letter (A, B, C, . . .). If exactly one such letter—or several copies of the same letter—appears, that letter is taken as the prediction.
2. When multiple different letters are detected, the code falls back to a full-phrase check: it normalises the text and accepts an answer only if one option's phrase appears while all others are absent.
3. If neither test succeeds (a situation that never arose with OLMo but did occur with Llama-2 65 B, which sometimes emits non-option tokens), no choice is selected and the response is automatically marked incorrect.

E Alternative Scoring settings

In addition to the scoring method presented in the main text (i.e., **ignore negated concepts**), we experimented with two additional scoring formulations:

1. **Affirmation/negation agnostic**: affirmation and negation labels of concepts are disregarded, so $C'(d, x_i) = C(d, x_i)$ for every candidate diagnosis–concept pair, and $k' = k$.
2. **Reward absence of negated concepts**: for a negated concept x_i we consider the number of times the candidate diagnosis d does *not* co-occur with x_i, so

$$C'(d, x_i) = \begin{cases} C(d, x_i) & \text{if } x_i \text{ is affirmed} \\ C(d) - C(d, x_i) & \text{if } x_i \text{ is negated,} \end{cases}$$

   with $k' = k$.

The three scoring methods yield highly consistent results on MEDQA diagnosis subset, largely due to the small number of negated concepts.

F Additional Results

F.1 RedPajama-based FBPR and LLaMA-65B alignment

For RedPajama-based FBPR and LLaMA-65B, their joint accuracy on is 24.2% on the MEDQA training diagnosis set; LLaMA alone was correct in 22.8% questions, while FBPR alone has an accuracy of 20.3%. They gave the same answer to 34.5% of the cases, among which 70.2% are correct.

F.2 Additional table and graphs

Corpus	k	all	Step 1	Step 2&3
Dolma	5	46.7%	51.2%	44.3%
Red Pajama	5	44.5%	46.1%	43.6%
Red Pajama	unrestricted	40.9%	40.2%	41.3%

Table 2: Accuracy of FBPR on MEDQA diagnosis subset (n=719), based on occurrence frequency from different corpora and with two settings of k, stratified on the medical licensing exam domain.

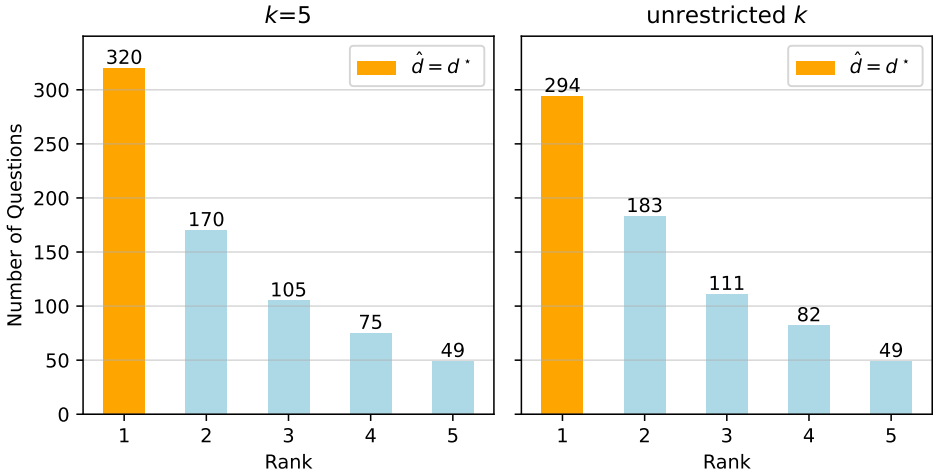

Figure 5: Distribution of the rank of the correct diagnosis of each question by RedPajama-based FBPR with $k = 5$ or unrestricted k.

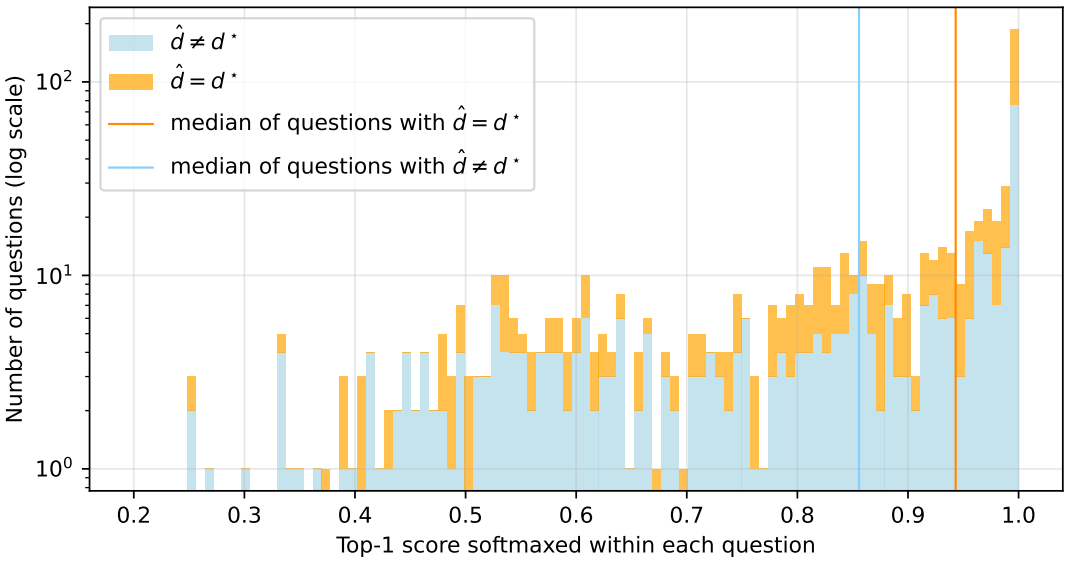

Figure 6: Stacked histogram of per-question softmaxed top-1 scores calculated by Dolma-based FBPR with $k = 5$. A larger softmaxed top-1 score indicates that FBPR's chosen diagnosis is much more strongly favored by frequency-based Naive Bayes relative to the other candidates.

F.3 FBPR pipeline examples

A success ($\hat{d} = d^\star$) and a failure ($\hat{d} \neq d^\star$) example of FBPR with are shown in Fig 7.

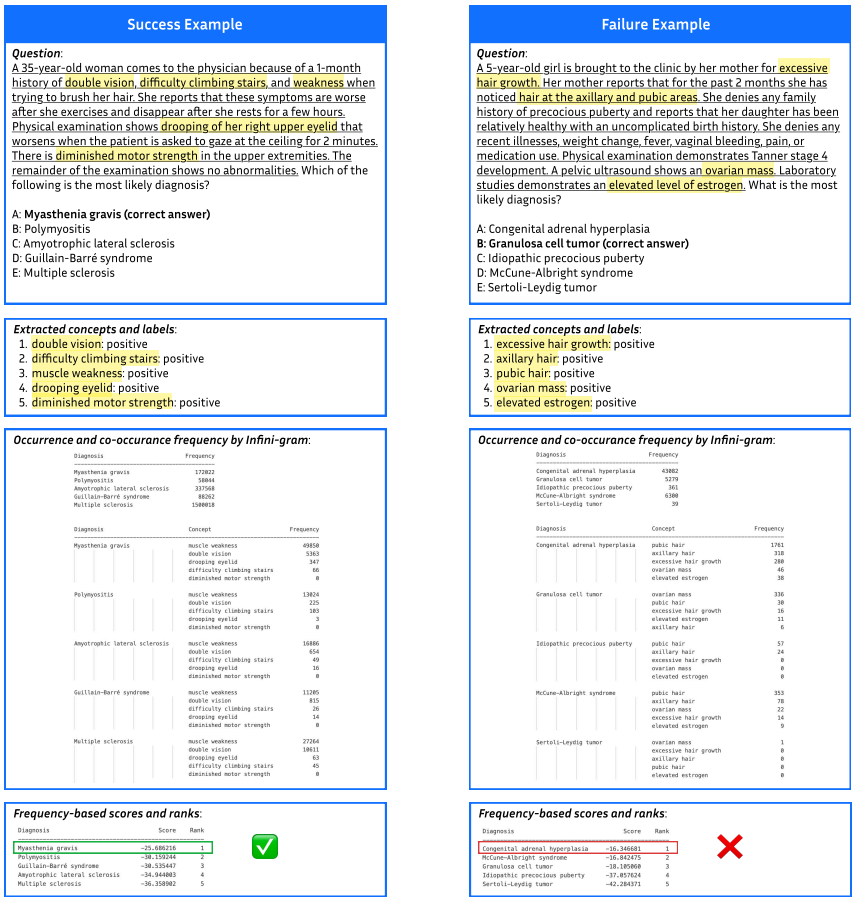

Figure 7: A success example (left) and a failure example (right) of FBPR on MEDQA diagnosis subset. The diagnosis occurrence and (diagnosis, concept) co-occurrence frequencies are retrieved by *Infini-gram* from `v4_rpj_llama_s4` index, i.e., from the Red Pajama corpus. By *Infini-gram* design, some of these frequencies are estimates.