# 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 that ask for the most likely diagnosis. We introduce 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 are sourced from the pretraining corpora for OLMo and Llama, FBPR achieves comparable performance to the corresponding LLMs pretrained on that same corpus. Furthermore, direct LLM inference and FBPR exhibit complementary strengths, correctly answering largely different subset of questions with overlap barely above chance. Combining them with a simple hybrid method yields a notable gain over either method alone. These findings show that explicit probabilistic baselines can account for a substantial portion of benchmark performance, while remaining valuable as meaningful, historically grounded reference points as well as complementary signals for hybridization.

## 1 Introduction

Large Language models (LLMs) have been used for tasks requiring probabilistic reasoning, including direct inference and estimation of priors (Paruchuri et al., 2024; Feng et al., 2025; 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., 2023; Singhal et al., 2023, 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 For example, expert systems such as DxPlain still surpass LLMs on unpublished diagnostic cases (Barnett et al., 1987; Berner et al., 1994; Feldman et al., 2025). This suggests a possible disconnect between performance on medical licensing exams and true probabilistic reasoning, with research highlighting LLM failures in metacognition (Griot et al., 2025a), processing new information under uncertainty (McCoy et al., 2025), and aligning with physicians' reasoning in 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-designed, data-refinable Bayesian network to provide probabilistic, patient-specific decision support. Recent evidence suggests that such systems can still surpass LLMs on unpublished diagnostic cases

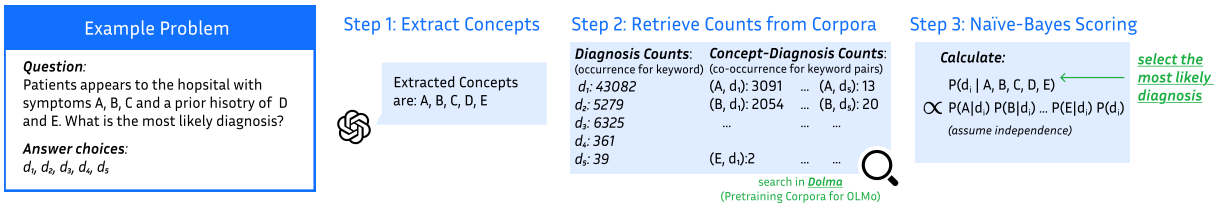


Figure 1: Frequency-Based Probabilistic Ranker: (1) concept extraction from the question, (2) corpus frequency retrieval, and (3) calculate a Naive-Bayes scoring to select the most likely diagnosis.

(Nee and Hein, 2010; Elkin et al., 2010; Bauer et al., 2002; Feldman et al., 2025).

Recent work uses LLM-derived probabilities for Bayesian reasoning, including Bayesian-network inference and parameterization (Feng et al., 2025; Nafar et al., 2025). In contrast, we ask how much a coarser, low-complexity, and transparent, frequency-based method can achieve on diagnosis questions using only an LLM's pretraining corpus. Therefore, we introduce **Frequency-Based Probabilistic Ranker (FBPR)** that eschews complex reasoning (Figure 1), and 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, which 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-tuned LLM pretrained on the Dolma corpus, and LLaMA 65B (Touvron et al., 2023a), 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%). A simple hybrid method combining FBPR with LLMs further improves performance (47.3% and 50.8%) across both corpora, highlighting the complementarity of the two approaches and the potential of using them in combination.

## 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 instruction-tuned. 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 $\leq 4$ 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" features 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. GPT-4o is used only for extraction and polarity labeling, not for scoring answer choices. Prompts and examples are detailed in Appendix B and Appendix H.2.

**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** To isolate corpus-frequency signal rather than model all clinical dependencies, we assume concept independence and conditional independence given each candidate diagnosis, 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 $\mathbf{x} = \{x_1, \ldots, x_k\}$.

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

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

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

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

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% |
| Hybrid (OLMo + Dolma) | 47.3% |
| FBPR on RedPajama | 44.5% |
| LLaMA 65B | 47.0% |
| Hybrid (LLaMA 65B + RedPajama) | 50.8% |

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 reached 46.7% accuracy and the RedPajama-based one 44.5% (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 shows 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 6 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.

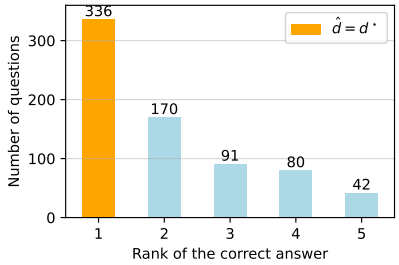


Figure 2: Distribution of the rank of the correct diagnosis of each question by Dolma-based FBPR with $k = 5$.

**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 6 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 RedPajama-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, indicating only a mild correlation. FBPR alone is correct on 21.9% of questions, and OLMo alone on 19.3%.

Across the diagnostic subset, the two methods agree on 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. This pattern also holds for RedPajama-based FBPR and LLaMA 65B (Appendix G.1).

**A simple hybrid method.** To leverage complementarity, our straightforward hybrid method combines the two signals by outputting the LLM's letter answer if it falls within FBPR's top $K$ ranked options; otherwise output FBPR's top-1. We fix $K = \lceil |D|/2 \rceil = 3$ in the five-option setting (full $K$ sweep in Appendix G.3). As Table 1 shows, this method shows gains on both signals on both corpus-based pairs.

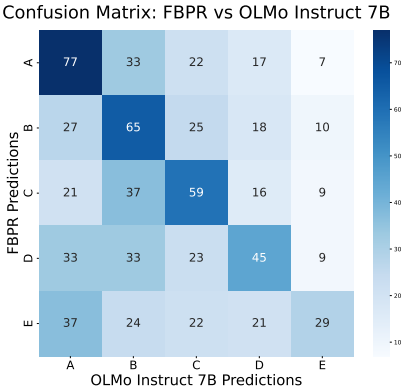


Figure 3: Confusion matrix between FBPR on Dolma and OLMo Instruct 7B on MEDQA diagnosis subset.

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 pretraining corpora statistics accounts for a substantial portion of benchmark performance. As the two approaches' signals are complementary, the two can be combined into a stronger predictor.

## 4 Discussion

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 that combine constructively, as our simple hybrid demonstrates (§3.2).

The strength of this signal is particularly notable given the minimal preprocessing in our implementation. FBPR performs no synonym or semantic normalization (e.g., "high blood pressure," "elevated blood pressure," and "hypertension" remain distinct concepts). While incorporating richer concept normalization could further improve FBPR performance (see Appendix F for illustrative cases), the simplicity of our method actually underscores how even a very naive baseline can be competitive. Additionally, future work could explore additional methods of probing LLMs to distinguish corpus-based pattern retrieval from more structured forms of clinical probabilistic 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. 2025. Bird: A trustworthy bayesian inference framework for large language models. In *International Conference on Learning Representations*, volume 2025, pages 8961–8989.

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. *Preprint*, arXiv:2505.24040.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

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. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):1–12.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*.

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. In Ger Devlin, editor, *Decision Support Systems - Advances in*, chapter 9. IntechOpen, London.

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, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

Akshay Paruchuri, Jake Garrison, Shun Liao, John B Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. 2024. What are the odds? language models are capable of probabilistic reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11712–11733, Miami, Florida, USA. Association for Computational Linguistics.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):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, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open

and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Maurice Weber, Daniel Y Fu, Quentin Gregory 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 Re, Irina Rish, and Ce Zhang. 2024. Redpajama: an open dataset for training large language models. In *The Thirty-eight Conference on Neural Information Processing Systems 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

We use GPT-4o only to extract surface clinical concepts from the vignette; the subsequent ranking stage does not use LLM inference over answer choices and relies only on corpus counts. 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 H.2.

### 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).

### B.3 Concept extraction

We use GPT-4o for phrase-level concept extraction because our downstream corpus query relies on exact-match co-occurrence counts. Off-the-shelf clinical entity extractors are often optimized for ontology mapping and may fragment coherent clinical phrases into generic entities, which are poorly matched to this retrieval setting. Examples comparing entities extracted using Google Healthcare API and phrase-level concepts from GPT-4o are demonstrated in Table 2. For example, in one MedQA vignette, GPT-4o extracts phrase-level concepts such as "difficulty climbing stairs" and "weakness brushing hair," while the Google Healthcare API extracts more fragmented entities such as "weakness" and "hair." Searching for co-occurrences with generic fragments such as "hair" can introduce substantial noise because the clinical context has been lost.

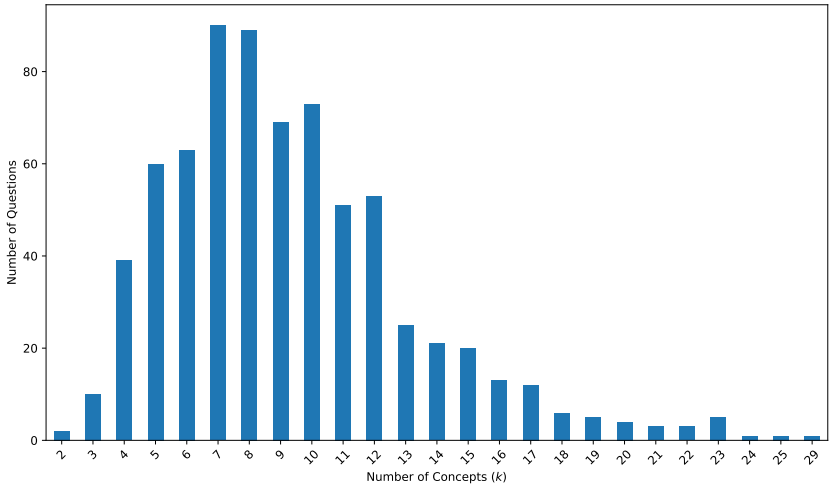


Figure 4: Distribution of the number of concepts per question extracted by GPT-4o when prompt contains no restriction.

| Text summary | GPT-4o phrase-level concepts | Google Healthcare API entities |
|---|---|---|
| Patient with double vision, difficulty climbing stairs, and weakness when brushing hair | double vision; difficulty climbing stairs; weakness brushing hair; symptoms worse after exercise; symptoms improve with rest; drooping right upper eyelid | double vision; weakness; hair; symptoms; worse; drooping; eyelid; diminished; motor strength; upper extremities |
| Patient with Crohn disease on natalizumab with difficulty walking and ataxia | crohn disease; difficulty walking; gait ataxia; limb ataxia; natalizumab | Crohn disease; difficulty walking; Crohn disease; natalizumab; intestinal symptoms; severe; unresponsive; therapies; gait; limb; ataxia |

Table 2: Examples illustrating why phrase-level extraction is better matched to exact-match co-occurrence retrieval than fragmented clinical entity extraction.

## 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, 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 $\mu m^3$
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 the LLM response. In the example above, the extracted answer refers to option choice B. We adapt a method similar to Jeong et al. (2024) and apply five tiers in order:

1. **Stand-alone letter regex.** A strict regex finds exact, stand-alone option letters (A–E). If exactly one letter—or several copies of the same letter—appears, it is the prediction.
2. **LaTeX/Markdown emphasised letter.** Patterns such as `\strong{C}`, `**B**`, `__C__` are recognised; the wrapped letter is the prediction.
3. **Numbered-list dominant letter.** Patterns of the form `"1.B 2.B 3.B ..."` are parsed; if a single letter dominates the list (count $\geq 2$), it is the prediction.
4. **Single-option phrase match.** The text is normalised; a prediction is made only if one option's phrase appears while all others are absent.
5. **Most-frequent stand-alone letter.** If multiple distinct letters appear without consensus from tiers 1–4, the most-frequent stand-alone letter is the prediction.

## 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 Synonym normalization

FBPR uses exact matching between extracted concepts and answer options and applies no synonym expansion or semantic normalization. This appendix illustrates the practical implications of the choice through three representative cases. We focus on cases where the correct answer has zero co-occurrence with at least one extracted concept: in such cases an extracted concept contributes no signal to the correct option, while synonym expansion could in principle introduce signal. For each such concept, we obtain three synonym candidates from GPT-4o (temperature = 0) and use infini-gram with inclusion–exclusion to compute the exact disjunctive count of the original concept together with its three synonyms against each answer option; scores are then recomputed with these expanded counts.

**Example 1: Lack of synonyms negatively impacts the score.** For the question below, the extracted concept *throbbing headache* has zero co-occurrences with the correct answer *Hypertensive emergency* under exact matching, while the synonym set {*pulsatile headache, pounding headache, vascular headache*} is able to retrieve co-occurrences. Including these forms is sufficient to move *Hypertensive emergency* from rank 4 to rank 1.

> **Question:** A 72-year-old woman comes to the emergency department 4 hours after the sudden onset of a diffuse, dull, throbbing headache. During this time, she also reports blurred vision, nausea, and one episode of vomiting. She has a history of hypertension and type 2 diabetes mellitus. Her medications include hydrochlorothiazide, lisinopril, atorvastatin, and metformin. She has smoked 1 pack of cigarettes daily for the past 30 years and drinks 1–2 glasses of wine per day. Her temperature is 36.6 °C (97.9 °F), pulse is 90/min, respirations are 14/min, and blood pressure is 185/110 mm Hg. Fundoscopic examination shows bilateral blurring of the optic disc margins. Physical and neurologic examinations show no other abnormalities. A complete blood count and serum concentrations of electrolytes, urea nitrogen, and creatinine are within the reference range. Urinalysis shows 2+ protein but no WBCs or RBCs. Which of the following is the most likely diagnosis?
> **Options:**
> A: Ischemic stroke
> B: Subarachnoid hemorrhage
> C: Idiopathic intracranial hypertension
> D: Hypertensive emergency
> E: Transient ischemic attack

**Example 2: canonical synonyms do not necessarily change the score.** For the question below, the extracted concept *hypertension* has a clear near-canonical paraphrase in *high blood pressure*, along with variants such as *arterial hypertension* and *elevated blood pressure*. Adding these forms does not change the rank of the correct answer: the correct option continues to receive zero co-occurrence with all three synonyms, and the score gap to competitors widens as other options obtain higher co-occurences count with the synonym expansion.

**Question:** A 65-year-old man is brought into the emergency department by his wife for slurred speech and right-sided weakness. The patient has a significant past medical history of hypertension and hyperlipidemia. The wife reports her husband went to bed last night normally but woke up this morning with the symptoms mentioned. Physical examination shows right-sided hemiparesis along with the loss of vibration and proprioception. Cranial nerve examination shows a deviated tongue to the left. What is the most likely diagnosis?
**Options:**
A: Medial pontine syndrome
B: Lateral pontine syndrome
C: Dejerine syndrome
D: Wallenberg syndrome
E: Weber syndrome

**Example 3: some concepts have no canonical normalization target.** For the question below, the extracted concept *burnt rubber smell* has zero co-occurrences across all five options. Unlike the clean synonym pair *hypertension* / *high blood pressure*, this concept has no canonical medical paraphrase. Normalization here forces an arbitrary choice between lay-description preservation and information-discarding abstraction, with no surface-form rule to select between them.

**Question:** A 50-year-old man presents to his primary care provider complaining of smelling abnormal odors on several occasions. He says that he smells burnt rubber even though there is nothing burning and no one around him can smell what he does. This symptom has been intermittently bothering him for the past 6 months. Also during this period, he had occasional nosebleeds. He works as a high school teacher. Although his work gets a little stressful around the exam season, he says he is able to cope well. Family history is unremarkable. He does not smoke or drink alcohol and denies the use of any medication. Physical examination reveals unilateral nasal obstruction with some dried blood in the nasal passage. What is the most likely diagnosis?
**Options:**
A: Psychomotor epilepsy
B: Neuroblastoma
C: Hypnagogic hallucination
D: Substance abuse
E: Schizophrenia

These examples suggest that synonym normalization is potentially useful but not mechanically straightforward. Expansions to synonyms can recover missing signal in some cases, yet it may leave the correct option unchanged or even strengthen competing options in others. Moreover, some clinical descriptions lack a clear canonical synonym. We therefore view semantic normalization as a promising future extension rather than a simple pre-processing step. At the same time, the fact that FBPR remains competitive under this deliberately minimal and naive exact-match setup suggests that raw corpus-level frequency signals already capture substantial diagnostic information.

# G Corpus-model alignment and hybrid analyses

## G.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.

### G.2 Hybrid method variants

We compare three hybrid rules that combine the LLM regex letter with FBPR's per-option log-scores $\{S_d\}_{d \in D}$. Let

$$c = \max_{d \in D} \frac{\exp S_d}{\sum_{d' \in D} \exp S_{d'}}$$

denote the within-question softmax of FBPR's top-1 score (its distribution on the Dolma pair is shown in Figure 6): a value near 1 indicates FBPR places almost all probability mass on a single option (high certainty); a value near $1/|D| = 0.2$ indicates the five options score similarly (no useful signal).

**Rank gate (main, §3.2).** We use FBPR's corpus signal as coarse-grained filtering. We adopt the LLM's prediction only when FBPR deems it reasonable. Specifically, if the LLM's letter falls within FBPR's top-$K$ ranked options, we choose the LLM's output. Otherwise, FBPR vetoes the LLM and we fall back to FBPR's top-1 choice. The cutoff $K$ can be any integer in $\{1, \ldots, |D|\}$; the main paper (§3.2) reports results at $K = \lceil |D|/2 \rceil = 3$ (the median rank for $|D| = 5$), a choice that requires no hyperparameter tuning. A full sweep over $K \in \{1, \ldots, 5\}$ is reported in §G.3.

**Confidence gate.** Return FBPR's top-1 when $c > \tau$; otherwise return the LLM's letter. Trust FBPR only when its softmax mass concentrates on a single option; default to the LLM when FBPR is uncertain. Introduces one threshold $\tau$.

**Combined gate.** Return FBPR's top-1 when $c > \tau$; otherwise, apply the rank gate. Combines the strengths of the two: trust FBPR unconditionally when it is highly confident; fall back to the rank-veto logic everywhere else.

| | LLaMA-65B + RedPajama | | | OLMo-Instruct + Dolma | | |
|---|---|---|---|---|---|---|
| Method | Acc | $\Delta_{\text{LLM}}$ | $\Delta_{\text{FBPR}}$ | Acc | $\Delta_{\text{LLM}}$ | $\Delta_{\text{FBPR}}$ |
| LLM | 47.0 | — | — | 44.1 | — | — |
| FBPR alone | 44.5 | — | — | 46.7 | — | — |
| **Rank gate (main)** | **50.8** | +3.7 | +6.3 | 47.3 | +3.2 | +0.6 |
| Confidence gate, $\tau$=0.99 | 48.4 | +1.4 | +3.9 | 49.2 | +5.2 | +2.5 |
| Confidence gate, $\tau$=0.95 | 47.4 | +0.4 | +2.9 | 48.8 | +4.7 | +2.1 |
| Confidence gate, $\tau$=0.90 | 47.6 | +0.6 | +3.1 | 48.8 | +4.7 | +2.1 |
| Combined gate, $\tau$=0.99 | 49.5 | +2.5 | +5.0 | **50.1** | +6.0 | +3.4 |
| Combined gate, $\tau$=0.95 | 48.4 | +1.4 | +3.9 | 49.5 | +5.4 | +2.8 |
| Combined gate, $\tau$=0.90 | 48.3 | +1.3 | +3.8 | 49.4 | +5.3 | +2.7 |

Table 3: Accuracy (%) and accuracy change relative to the LLM baseline ($\Delta_{\text{LLM}}$) and to FBPR-alone baseline ($\Delta_{\text{FBPR}}$) for the three hybrid variants on both corpus-model pairs at three confidence thresholds $\tau$. Bold accuracy marks each pair's strongest configuration.

**Analysis.** The best variant depends on which baseline is stronger a priori. On LLaMA-65B + RedPajama (LLM > FBPR), the rank gate is strongest (50.8%): it effectively uses FBPR as a coarse-level filter while relying on the LLM for fine-grained selection. On OLMo + Dolma (FBPR > LLM), trusting FBPR's confidence as a stronger primary signal is superior, making the combined gate at $\tau = 0.99$ the best (50.1%). Accuracy remains stable across $\tau \in \{0.90, 0.95, 0.99\}$. We retain the rank gate as our main result because it is simple and yields consistent gains across both pairs.

### G.3 Rank-gate cutoff $K$ ablation

The rank-gate cutoff $K$ controls how lenient FBPR is with the LLM: the LLM letter is kept if it falls within FBPR's top-$K$ ranked options. The main paper uses $K = \lceil |D|/2 \rceil = 3$ in the five-option setting. Table 4 reports the rank gate at every $K \in \{1, 2, 3, 4, 5\}$. $K = 1$ keeps the LLM only when it agrees with FBPR's top-1, so the rule degenerates to FBPR alone. $K = 5$ keeps the LLM unconditionally (since it always lies within FBPR's top-5), so the rule degenerates to LLM alone, except on the 32 questions where the LLaMA 65B model returns *None* to be extracted by the regex rules, so they fall back to FBPR top-1.

In general, all $K$ values for meaningful hybrid ($K \in \{2, 3, 4\}$) would lead to improvement over the single method.

| | LLaMA-65B + RedPajama | | | OLMo-Instruct + Dolma | | |
|---|---|---|---|---|---|---|
| $K$ | Acc | $\Delta_{\text{LLM}}$ | $\Delta_{\text{FBPR}}$ | Acc | $\Delta_{\text{LLM}}$ | $\Delta_{\text{FBPR}}$ |
| 1 | 44.5 | $-2.5$ | 0.0 | 46.7 | $+2.6$ | 0.0 |
| 2 | 48.3 | $+1.3$ | $+3.8$ | **49.4** | $+5.3$ | $+2.6$ |
| **3 (main)** | **50.8** | $+3.7$ | $+6.3$ | 47.3 | $+3.2$ | $+0.6$ |
| 4 | 50.5 | $+3.5$ | $+6.0$ | 46.7 | $+2.6$ | 0.0 |
| 5 | 48.8 | $+1.8$ | $+4.3$ | 44.1 | 0.0 | $-2.6$ |

Table 4: Rank-gate accuracy (%) and change relative to the LLM baseline ($\Delta_{\text{LLM}}$) and FBPR-alone baseline ($\Delta_{\text{FBPR}}$) for $K \in \{1, \ldots, 5\}$. $K = \lceil |D|/2 \rceil = 3$ (paper main) is the peak on LLaMA + RedPajama; $K = 2$ is the peak on OLMo + Dolma. The corpus-dependent best $K$ mirrors the direction asymmetry of the confidence threshold $\tau$: when the LLM is the stronger baseline, the larger $K$ wins; when FBPR is stronger, the smaller $K$ wins. The boundary cases $K = 1$ (= FBPR alone, $\Delta_{\text{FBPR}} = 0$) and $K = 5$ ($\approx$ LLM alone, $\Delta_{\text{LLM}} \approx 0$) provide reference.

### G.4 Why combining helps on one pair but not the other: a bin-stratified view

To understand mechanistically how the combined gate's accuracy differs from its two component rules, we stratify the 719 questions into the two bins induced by FBPR's certainty $c$ at $\tau = 0.99$. In each bin, the combined gate is deterministically identical to one component:

- When $c > \tau$ (high certainty): combined gate = confidence gate = FBPR's top-1.
- When $c \leq \tau$ (low certainty): combined gate = rank gate.

Table 5 reports each method's accuracy within each bin.

| | LLaMA-65B + RedPajama | | OLMo + Dolma | |
|---|---|---|---|---|
| Method | $c > \tau$ ($n$=228) | $c \leq \tau$ ($n$=491) | $c > \tau$ ($n$=197) | $c \leq \tau$ ($n$=522) |
| LLM regex | 49.6 | 45.8 | 40.6 | 45.4 |
| FBPR alone | 50.0 | 42.0 | 59.4 | 42.0 |
| Rank gate | **54.0** | **49.3** | 49.2 | **46.6** |
| Confidence gate | 50.0 | 47.7 | **59.4** | 45.4 |
| Combined gate | 50.0 | **49.3** | **59.4** | **46.6** |

Table 5: Per-bin accuracy (%) at $\tau = 0.99$. By construction, the combined gate matches the confidence gate in the $c > \tau$ bin and the rank gate in the $c \leq \tau$ bin. Bold marks the strongest method per bin/pair.

**Corpus dependency of the combined gate.** The relative success of the combined gate is determined by its performance across two confidence bins. In the high-certainty bin ($c > \tau$), strictly adopting FBPR's top-1 prediction improves accuracy on OLMo but degrades it on LLaMA compared to the rank gate. In the low-certainty bin ($c \leq \tau$), however, the rank gate uniformly outperforms the confidence gate on both pairs, as its fallback mechanism recovers correct predictions even when FBPR is uncertain. Consequently, the combined gate achieves an overall gain on OLMo where both conditions align, but falls slightly behind the rank gate on LLaMA due to the high-certainty degradation. This demonstrates that the combined gate's effectiveness is an inherent property of the corpus, specifically the reliability of FBPR's highly confident predictions.

## H Additional FBPR results and examples

### H.1 Unrestricted-$k$ and certainty results

| Corpus | $k$ | all | Step 1 | Step 2&3 |
|---|---|---|---|---|
| Dolma | 5 | 46.7% | 51.2% | 44.3% |
| Dolma | unrestricted | 45.1% | 47.7% | 43.6% |
| Red Pajama | 5 | 44.5% | 46.1% | 43.6% |
| Red Pajama | unrestricted | 40.9% | 40.2% | 41.3% |

Table 6: 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.

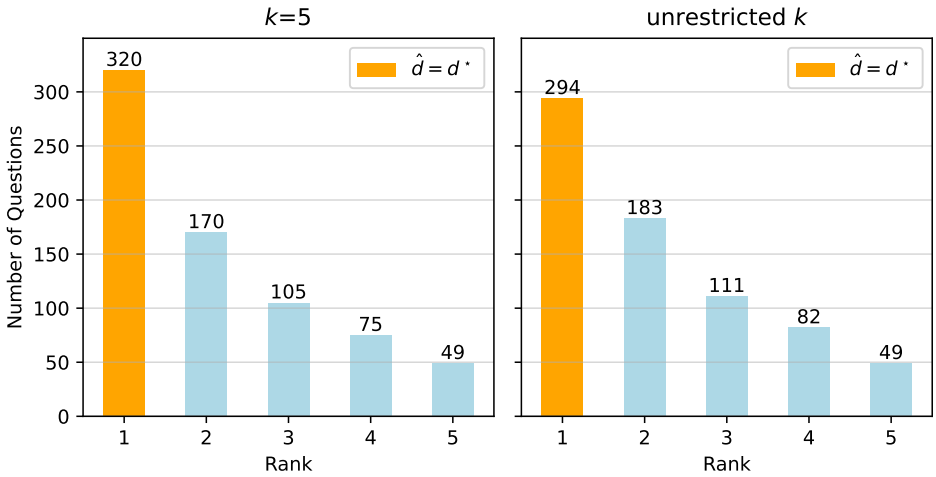


Figure 5: Distribution of the rank of the correct diagnosis of each question by RedPajama-based FBPR with $k = 5$ or unrestricted $k$.

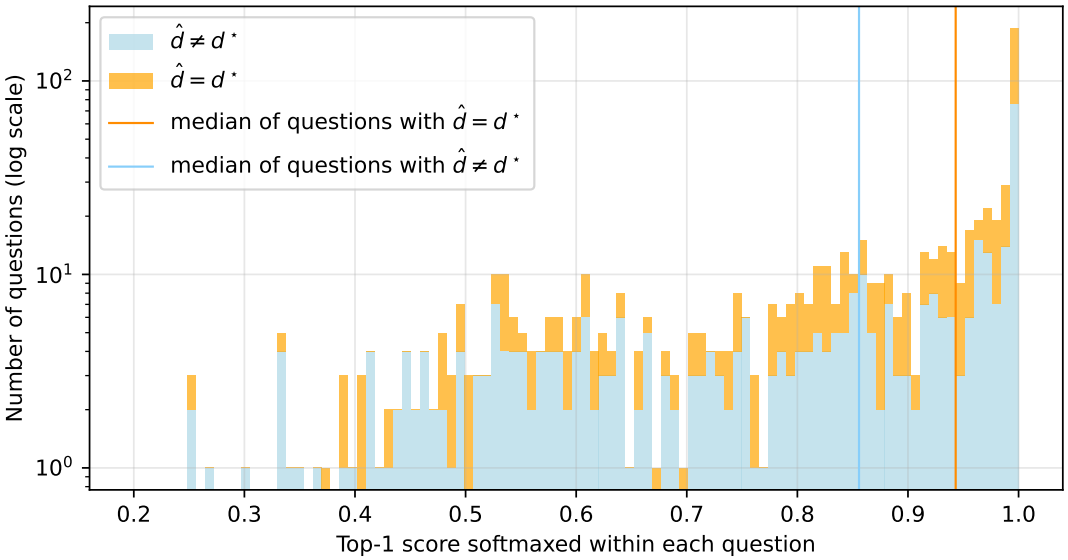


Figure 6: Stacked histogram of per-question softmaxed top-1 scores calculated by Dolma-based FBPR with $k = 5$. A larger softmaxed top-1 score indicates that FBPR's chosen diagnosis is much more strongly favored by frequency-based Naive Bayes relative to the other candidates.

## H.2 FBPR pipeline examples

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

**Success Example**

***Question***:
A 35-year-old woman comes to the physician because of a 1-month history of double vision, difficulty climbing stairs, and weakness when trying to brush her hair. She reports that these symptoms are worse after she exercises and disappear after she rests for a few hours. Physical examination shows drooping of her right upper eyelid that worsens when the patient is asked to gaze at the ceiling for 2 minutes. There is diminished motor strength in the upper extremities. The remainder of the examination shows no abnormalities. Which of the following is the most likely diagnosis?

A: **Myasthenia gravis (correct answer)**
B: Polymyositis
C: Amyotrophic lateral sclerosis
D: Guillain-Barré syndrome
E: Multiple sclerosis

***Extracted concepts and labels***:
1. double vision: positive
2. difficulty climbing stairs: positive
3. muscle weakness: positive
4. drooping eyelid: positive
5. diminished motor strength: positive

***Occurrence and co-occurance frequency by Infini-gram***:

| Diagnosis | Frequency |
|---|---|
| Myasthenia gravis | 172022 |
| Polymyositis | 58044 |
| Amyotrophic lateral sclerosis | 337568 |
| Guillain-Barré syndrome | 88262 |
| Multiple sclerosis | 1500018 |

| Diagnosis | Concept | Frequency |
|---|---|---|
| Myasthenia gravis | muscle weakness | 49850 |
| | double vision | 5363 |
| | drooping eyelid | 347 |
| | difficulty climbing stairs | 66 |
| | diminished motor strength | 0 |
| Polymyositis | muscle weakness | 13024 |
| | double vision | 225 |
| | difficulty climbing stairs | 103 |
| | drooping eyelid | 3 |
| | diminished motor strength | 0 |
| Amyotrophic lateral sclerosis | muscle weakness | 16886 |
| | double vision | 654 |
| | difficulty climbing stairs | 49 |
| | drooping eyelid | 16 |
| | diminished motor strength | 0 |
| Guillain-Barré syndrome | muscle weakness | 11205 |
| | double vision | 815 |
| | difficulty climbing stairs | 26 |
| | drooping eyelid | 14 |
| | diminished motor strength | 0 |
| Multiple sclerosis | muscle weakness | 27264 |
| | double vision | 10611 |
| | drooping eyelid | 63 |
| | difficulty climbing stairs | 45 |
| | diminished motor strength | 0 |

***Frequency-based scores and ranks***:

| Diagnosis | Score | Rank |
|---|---|---|
| Myasthenia gravis | −25.686216 | 1 |
| Polymyositis | −30.159244 | 2 |
| Guillain-Barré syndrome | −30.535447 | 3 |
| Amyotrophic lateral sclerosis | −34.944003 | 4 |
| Multiple sclerosis | −36.358902 | 5 |

✅

**Failure Example**

***Question***:
A 5-year-old girl is brought to the clinic by her mother for excessive hair growth. Her mother reports that for the past 2 months she has noticed hair at the axillary and pubic areas. She denies any family history of precocious puberty and reports that her daughter has been relatively healthy with an uncomplicated birth history. She denies any recent illnesses, weight change, fever, vaginal bleeding, pain, or medication use. Physical examination demonstrates Tanner stage 4 development. A pelvic ultrasound shows an ovarian mass. Laboratory studies demonstrates an elevated level of estrogen. What is the most likely diagnosis?

A: Congenital adrenal hyperplasia
**B: Granulosa cell tumor (correct answer)**
C: Idiopathic precocious puberty
D: McCune-Albright syndrome
E: Sertoli-Leydig tumor

***Extracted concepts and labels***:
1. excessive hair growth: positive
2. axillary hair: positive
3. pubic hair: positive
4. ovarian mass: positive
5. elevated estrogen: positive

***Occurrence and co-occurance frequency by Infini-gram***:

| Diagnosis | Frequency |
|---|---|
| Congenital adrenal hyperplasia | 43082 |
| Granulosa cell tumor | 5279 |
| Idiopathic precocious puberty | 361 |
| McCune-Albright syndrome | 6300 |
| Sertoli-Leydig tumor | 39 |

| Diagnosis | Concept | Frequency |
|---|---|---|
| Congenital adrenal hyperplasia | pubic hair | 1761 |
| | axillary hair | 318 |
| | excessive hair growth | 280 |
| | ovarian mass | 46 |
| | elevated estrogen | 38 |
| Granulosa cell tumor | ovarian mass | 336 |
| | pubic hair | 30 |
| | excessive hair growth | 16 |
| | elevated estrogen | 11 |
| | axillary hair | 6 |
| Idiopathic precocious puberty | pubic hair | 57 |
| | axillary hair | 24 |
| | excessive hair growth | 0 |
| | ovarian mass | 0 |
| | elevated estrogen | 0 |
| McCune-Albright syndrome | pubic hair | 353 |
| | axillary hair | 78 |
| | ovarian mass | 22 |
| | excessive hair growth | 14 |
| | elevated estrogen | 9 |
| Sertoli-Leydig tumor | ovarian mass | 1 |
| | excessive hair growth | 0 |
| | axillary hair | 0 |
| | pubic hair | 0 |
| | elevated estrogen | 0 |

***Frequency-based scores and ranks***:

| Diagnosis | Score | Rank |
|---|---|---|
| Congenital adrenal hyperplasia | −16.346681 | 1 |
| McCune-Albright syndrome | −16.842475 | 2 |
| Granulosa cell tumor | −18.105060 | 3 |
| Idiopathic precocious puberty | −37.057624 | 4 |
| Sertoli-Leydig tumor | −42.284371 | 5 |

❌

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.